National and International Educational Assessments: Overview, Results, and Issues

Updated November 2, 2018

Congressional Research Service
https://crsreports.congress.gov
R45401

CRS REPORT
Prepared for Members and
Committees of Congress

Congressional Research Service
Informing the legislative debate since 1914

SUMMARY

R45401

November 2, 2018

Rebecca R. Skinner
Specialist in Education Policy

National and International Educational Assessments: Overview, Results, and Issues

U.S. students participate in many assessments to track their educational achievement. Perhaps the most widely discussed of these are statewide assessments required by the Elementary and Secondary Education Act (ESEA), which was most recently comprehensively amended by the Every Student Succeeds Act (ESSA; P.L. 114-95). However, U.S. students also participate in large-scale national assessments, authorized by the National Assessment of Educational Progress Assessment Act (NAEPAA; Title III, Section 303 of P.L. 107-279), and international assessments, authorized by the Education Sciences Reform Act (ESRA; Title I, Section 153(a)(6) of P.L. 107-279). At the national level, students participate in the National Assessment of Educational Progress (NAEP). At the international level, U.S. students participate in the Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Program for International Student Assessment (PISA).

Although there are some similarities between statewide, national, and international assessments, they differ in purpose and level of reporting. For example, the purpose of statewide assessments is primarily to inform statewide accountability systems and provide information on individual achievement. By contrast, the purpose of large-scale assessments is to highlight achievement gaps, track national progress over time, compare achievement within the United States, and compare U.S. achievement to that of other countries. Results of these assessments are not reported for individuals.

National Assessments: The NAEP is a series of assessments measuring achievement in various content areas. The long-term trends NAEP (LTT NAEP) has tracked achievement since the 1970s and has remained relatively unchanged. The main NAEP assessment has tracked achievement since the 1990s and changes periodically to reflect changes in school curricula. The main NAEP has three levels: national, state, and Trial Urban District Assessment (TUDA). States that receive Title I-A funding under the ESEA are required to participate in biennial state NAEP assessments in reading and mathematics for 4th and 8th grade. Results from the 2017 main NAEP show a small but significant increase in 8th grade reading since 2015. There were no significant changes in 4th grade reading, 4th grade mathematics, or 8th grade mathematics since 2015. Longer term, however, average reading and mathematics scores have increased significantly since the initial administrations in the 1990s.

International Assessments: The United States participates in three international assessments: TIMSS, PIRLS, and PISA. TIMSS is an assessment of mathematics and science for 8th grade students. PIRLS is an assessment of reading literacy for 4th grade students. PISA is an assessment of reading literacy, mathematics literacy, and science literacy for 15-year-old students. In general, U.S. students have made statistically significant gains since the initial administrations of international assessments; however, achievement did not consistently increase in the most recent administrations of international assessments.

Issues of Interpretation of National and International Assessments: Results of national and international assessments are difficult to interpret. One challenge is processing the large amount of data. Another is understanding the difference between statistical significance and educational significance. Reporting statistical significance is standard practice in research, but it does not convey the magnitude of a difference and its associated educational significance. Another issue is the tendency to focus narrowly on one assessment at one point in time. A narrow focus may not provide the appropriate context to interpret results accurately. International assessment results may also be affected by socioeconomic considerations within and across countries.

Comparing Results Across Assessments: Comparing results across national and international assessments can be challenging. Each assessment was created for a unique purpose by different groups of stakeholders, which makes direct comparisons difficult. There are a number of issues to consider when evaluating U.S. students' performance across assessments. For example, consideration must be given to the differences in (1) the degree of alignment of content standards and assessments, (2) the target population being assessed, (3) the voluntary nature of student participation, (4) the participating education systems, (5) the scale of the assessment, and (6) the precision of measurement for each assessment.

Contents

Figures

Tables

Appendixes

Contacts

Overview

Assessing the achievement of students in elementary and secondary schools and the nation's educational progress is fundamental to informing education policy approaches. Congressional interest in this area includes and extends beyond the annual assessments administered by states to comply with the educational accountability requirements of Title I-A of the Elementary and Secondary Education Act (ESEA). Congressional interest in testing also encompasses a national assessment program, authorized by the National Assessment of Educational Progress Assessment Act (NAEPAA; Title III, Section 303 of P.L. 107-279), and participation in international assessment programs, authorized by the Education Sciences Reform Act (ESRA; P.L. 107-279, Section 153(a)(6)). At the national level, students participate in the National Assessment of Educational Progress (NAEP). At the international level, U.S. students participate in the Trends in International Mathematics and Science Study (TIMSS), Progress in International Reading Literacy Study (PIRLS), and Program for International Student Assessment (PISA).[1]

When national and international assessment results are released, there is a tendency to take the results of one assessment and present them as a snapshot of U.S. student achievement. The focus on one set of assessment outcomes may result in a narrow and possibly misleading view of overall student achievement. The primary purpose of this report is to provide background and context for the interpretation of national and international assessment scores so that results can be interpreted appropriately over time and across multiple assessments. Other purposes of this report are to describe specific national and international assessments, describe the recent results of these assessments, and clarify specific issues regarding the interpretation of assessment scores that explain the achievement of U.S. students.

Introduction to Large-Scale Assessments

National and international assessments are large-scale assessments of educational progress. While some may also consider statewide assessments "large-scale," for the purposes of this report "large-scale assessments" refers only to national and international assessments. These assessments differ from statewide and other assessments in several important ways. First, large-scale assessments have different purposes than smaller-scale assessments. Second, there are different participation requirements and sampling procedures. And, third, there are differences in the ways scores are typically reported for large-scale assessments versus smaller-scale assessments. This section of the report discusses some of the major differences between large-scale and other, smaller-scale assessments, such as state and local assessments.

Purposes of Large-Scale Assessments

The primary purposes of large-scale assessments are to highlight achievement gaps, track national progress over time, compare student achievement within the United States, and compare U.S. academic performance to the performance of other countries. Unlike statewide assessments that evaluate schools and districts, large-scale assessment results generally cannot be connected to individual students, schools, or districts.[2] Results are typically reported at the national or state levels.

[1] See **Appendix C** for a glossary of the acronyms used in this report.

[2] It is possible to connect national assessment results to some large urban districts in the United States. The National

Results from large-scale assessment that are reported at the national or state levels are well suited for broad-based analyses of achievement gaps in the United States. The "achievement gap" refers to differences in educational performance across subgroups of U.S. students. The most commonly reported achievement gaps are those that highlight differences by race, ethnicity, socioeconomic status, disability status, and gender.

Results reported at the national and state level are also well suited to track U.S. progress over time. Statewide assessments tend to change periodically depending on several factors, including changes in state and federal legislative requirements, in the assessments administered, and in the vendors that assist states with assessment development. By contrast, national and international assessments have remained relatively stable over time. Due to this stability, the results are easier to interpret from year to year because they are more of a direct comparison. Some national assessment programs date back to the 1960s and allow for a broader view of educational progress than statewide assessments.

Results reported at the national level on international assessments are uniquely suited to compare U.S. academic performance to the performance of other countries. Statewide assessments and national assessments cannot be used for this purpose. The United States has participated in international assessments since the 1960s.[3] Depending on the type of international assessment and year of administration, U.S. student performance has been compared to student performance in approximately 30 to 70 countries. Furthermore, some international assessments have been benchmarked against U.S. student performance in certain states.[4]

Participation

Participation requirements for statewide, national, and international assessments differ. States are required by the ESEA to assess all students in statewide assessment programs, including students with disabilities and English Learners (ELs).[5] In assessment terminology, states are required to assess the "universe" of students (i.e., all students) in statewide assessments.

States that receive Title I-A ESEA funding (currently, all states) are also required to participate in biennial NAEP assessments of reading and mathematics for the 4th and 8th grades. In contrast to statewide assessments, however, states are required to administer national assessments to a subset of students. In assessment terminology, states assess a "representative sample" of students.[6] Additionally, states may not be required to administer the assessments to certain students with

Assessment of Educational Progress (NAEP) includes a program called the Trial Urban District Assessment (TUDA). The TUDA currently collects and reports information on 27 large urban school districts in the United States. For more information on TUDA, see https://nces.ed.gov/nationsreportcard/about/district.aspx. More commonly, however, results of NAEP are reported at the national and state levels.

[3] The United States participated in the First International Mathematics Study (FIMS) in 1964. FIMS was followed by the Second International Mathematics Study (SIMS) in 1990 and the Third International Mathematics and Science Study (TIMSS) in 1995. TIMSS was renamed to the Trends in International Mathematics and Science Study (also TIMSS) and continues to the present day.

[4] Comparisons of U.S. states to other countries will be covered in a later section, "Comparing Results Across Assessments."

[5] In practice, states are required to report assessment results for 95% of all students and 95% of all student groups. For more information, see CRS Report R45049, *Educational Assessment and the Elementary and Secondary Education Act*.

[6] A representative sample is a group that closely matches the characteristics of its population as a whole. For more information on selecting a representative sample in the national assessment, see https://nces.ed.gov/nationsreportcard/assessment_process/selection.aspx#samples.

disabilities and ELs if these students require an accommodation that is not permitted on the national assessments.[7] Although states are required to participate in these assessments, individual participation of students remains voluntary.

Unlike statewide assessments and the NAEP assessment, states are not required to participate in many national assessments and or any international assessments. Participation is voluntary at both the state and student levels. If a state agrees to participate, each international assessment has a different method for selecting students. Like national assessments, international assessments test a "representative sample" of students.

Score Reporting

Score reporting for large-scale and smaller-scale assessments has some noteworthy similarities and differences. Statewide, national, and international assessments can all report student achievement as scaled scores.[8] A scaled score is a standardized score that exists among a common scale that can be used to make comparisons across students, across subgroups of students, and over time on a given assessment.

Educational assessment often reports scaled scores instead of raw scores or percent correct. There are several reasons that scaled scores are preferable. Large-scale assessment programs usually have multiple forms of the same test to control for student exposure to assessment items.[9] As such, students take multiple forms of the same test. Although the multiple forms of the same assessment are similar, there are inevitably differences in difficulty of certain items across forms. By creating a scaled score, the scores of students or groups of students can be directly compared, even when different forms of varying difficulty were administered.[10]

Although all these assessments use scaled scores, they all have a different scale. For example, some scales from national assessments are from 0 to 300 while scales from international assessments are typically 0 to 1,000.[11] Therefore, scaled scores are not directly comparable. When a scaled score is reported in isolation, it may be difficult to determine how well a student or group of students performed. To provide a context for grade-level or age-level expectations, large-scale assessments (and some smaller-scale assessments, such as the statewide assessments required by Title I-A of the ESEA) use performance standards.

---

[7] An accommodation is a change in testing materials or procedures that allows students with disabilities or ELs to show their knowledge and skills. For general information about accommodations, see the National Center for Education Outcomes description at https://nceo.info/Assessments/general_assessment/accommodations. For specific information on the accommodations allowed on the NAEP, see https://nces.ed.gov/nationsreportcard/about/accom_table.aspx.

[8] For more information on scale scores, see CRS Report R45048, *Basic Concepts and Technical Considerations in Educational Assessment: A Primer*.

[9] Many standardized assessments control for exposure to assessment items. As part of the assessment program, a number of assessment items are released to the public as examples. In addition, students who have taken the assessment in the past have prior knowledge of certain assessment items. Due to this exposure, multiple forms of assessments are developed so that there are no practice effects.

[10] For more information, see Xuan Tan and Rochelle Michel, "Why Do Standardized Testing Programs Report Scaled Scores? Why Not Just Report the Raw or Percent-Correct Scores?," *ETS R&D Connections*, vol. 16 (September 2011), https://www.ets.org/Media/Research/pdf/RD_Connections16.pdf.

[11] The scale for an assessment represents the full range of achievement for all intended test takers. For example, one assessment can potentially have a vertical scale from 0 to 1000 that measures K-12 reading achievement. A student in third grade may score 300 on the scale and be "proficient" in third-grade reading. A student in fifth grade may score 300 on the scale and be "not proficient" in fifth-grade reading. The scale, therefore, interacts with the grade level and expectations for the student. There is no single number that denotes proficiency for all students.

A performance standard is an agreed-upon definition of a certain level of performance in a content area that is expressed in terms of a cut score (i.e., basic, proficient, advanced) for a specific assessment. Although statewide, national, and international assessments use performance standards and may even use the same terminology (e.g., basic, proficient, advanced) to describe one or more of their performance standards, they do not use the same performance standards. For each assessment, there may be different cut scores and different definitions of each performance level. A student who is "proficient" on a statewide assessment may not be "proficient" on a national assessment and vice versa. In addition, within an individual assessment, the range of actual student performance within the "proficient" performance standard, for example, will include students whose assessment results are just high enough to be considered proficient as well as students whose assessment results almost put them into the next-highest performance standard level (e.g., advanced). In this example, the "proficient" performance standard does not distinguish between a student who is just barely proficient and one who is nearly advanced. Both students would be considered to be proficient.

International assessments usually report scores differently than statewide and national assessments. Although international assessments do report a scaled score and sometimes a performance standard, they have additional ways of reporting achievement. Performance on international assessments is also reported as a rank or as a score relative to an "international average" score. Rank and international average scores tend to change from one assessment administration to the next, depending on the countries that participate in the assessment.

Perhaps the most important distinction between statewide and large-scale assessments is the level of reporting. Statewide assessments are administered so that scores can be reported for individual students. Because statewide assessment programs test the universe of students and each student takes all the assessment items, each student has his or her own scaled score and performance standard level (e.g., basic, proficient, advanced). In large-scale assessments, a representative sample of students is tested, and each student may only take a portion of the assessment items. This type of sampling procedure allows scores to be reported for groups of students but not individual students. Large-scale assessments, therefore, report scores for groups of students with similar demographic characteristics, groups within a large district or state, or groups within a country.

Types of Large-Scale Assessments

Large-scale assessments are standardized assessments that are administered nationwide or worldwide. U.S. students currently participate in two types of large-scale assessments: national assessments and international assessments.

The United States administers a series of national assessments called the National Assessment of Educational Progress. Although NAEP is described as a single assessment, it is actually a series of two assessment programs: the main NAEP and the long-term trends (LTT) NAEP. The main NAEP program consists of three subprograms: national NAEP, state NAEP, and the Trial Urban District Assessment (TUDA). The United States also participates in three major international assessments: the Trends in International Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Program for International Student Assessment (PISA).

Table 1 provides a quick reference guide to the characteristics of the large-scale assessments discussed in this report. **Appendix A** provides additional information on large-scale assessments, such as authorization and oversight provisions.

Table 1. Large-Scale Assessment Characteristics

Assessment Title	Content Areas	Grade Levels or Ages	Student Participation	Initial Assessment	Frequency of Administration
National Assessment of Educational Progress (NAEP)[a]					
National NAEP	Reading, mathematics, science, writing, the arts, civics, economics, geography, U.S. history, and technology and engineering literacy (TEL)[b]	4th grade, 8th grade, and 12th grade (less frequently)	Representative sample selected Voluntary participation for states and students	1969	Variable[c]
State NAEP	Reading, mathematics, science, and writing	4th grade, 8th grade, and 12th grade (less frequently)	Representative sample selected Required participation in 4th and 8th grade reading and mathematics assessments for states that receive ESEA, Title I-A funding Voluntary participation for students	1990	Every 2 years[c]
Trial Urban District Assessment (TUDA) NAEP	Reading, mathematics, science, and writing	4th grade, 8th grade, and 12th grade (less frequently)	Representative sample selected Voluntary participation for districts	2003[d]	Every 2 years
Long-Term Trends (LTT) NAEP	Reading and mathematics[e]	9-, 13-, and 17-year-olds	Representative sample selected Voluntary participation for states and students	1969	LTT NAEP is administered "regularly" but the frequency of administration has ranged from about every 2 to 12 years.

Assessment Title	Content Areas	Grade Levels or Ages	Student Participation	Initial Assessment	Frequency of Administration
International Assessments					
Program for International Student Assessment (PISA)	Reading, mathematics, and science literacy	15-year-olds	Representative sample selected Voluntary participation for countries and students	2000	Every 3 years
Program for International Reading Literacy Study (PIRLS)	Reading, school and teacher practices related to instruction, students' attitudes towards reading, and reading habits	4th grade	Representative sample selected Voluntary participation for countries and students	2001	Every 5 years
Trends in International Mathematics and Science Study (TIMSS)	Mathematics and science[f]	4th grade, 8th grade, and 12th grade	Representative sample selected Voluntary participation for countries and students	1995	Every 4 years

Source: CRS summary of national and international assessments, available from the U.S. Department of Education (ED).

a. The NAEP has two assessment programs: main NAEP and LTT NAEP. The main NAEP has three subprograms: national NAEP, state NAEP, and TUDA NAEP. The main NAEP subprograms have significant overlap. The LTT NAEP differs from the main NAEP in its origin, frequency, and content areas assessed. For more information, see https://nces.ed.gov/nationsreportcard/about/ltt_main_diff.aspx.

b. The national NAEP does not assess each content area at each administration. For more information on the content areas assessed by year, see the NAEP assessment schedule: https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx.

c. The state NAEP does not assess each content area at each administration. Although state NAEP is typically assessed every two years, it was assessed in both 2002 and 2003. At this time, the TUDA program was in a trial period, and the timing of assessments were being coordinated across programs. For more information, see the NAEP assessment schedule: https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx.

d. The TUDA does not assess each content area at each administration. The initial TUDA was administered in 2002 in the content areas of reading and writing. The first TUDA administration to assess both reading and mathematics was in 2003. TUDA reading and mathematics has been assessed every two years since 2003. For more information, see the NAEP assessment schedule: https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx.

e. Historically, the LTT NAEP assessed a wider variety of content areas; however, content areas other than mathematics and reading have not been assessed since 1999 because NAGBE changed its policy on the LTT NAEP. For more information, see the *National Assessment Governing Board Long-Term Trends Policy Statement*, adopted May 18, 2002, available at https://www.nagb.gov/content/nagb/assets/documents/policies/Long-term%20Trend.pdf.

f. Specific mathematics and science skills depend on the grade level assessed. For more information, see https://nces.ed.gov/timss/faq.asp#7.

National Assessments: The National Assessment of Educational Progress

The NAEP is referred to as the "Nation's Report Card" because it is the only nationally representative assessment of what America's students know and can do in various content areas.[12] The original NAEP program began in 1969 and the first assessment was administered in 1971. The National Assessment of Educational Progress Authorization Act authorizes the NAEP.[13] The Commissioner for the National Center for Education Statistics (NCES) in the U.S. Department of Education (ED) is responsible for the administration of the NAEP. The Secretary of Education appoints members to the National Assessment Governing Board (NAGB) to set the policy for NAEP administration. The Commissioner of NCES and the NAGB meet regularly to coordinate activities.

In the first two decades of NAEP administration, there was no "main NAEP" program or "LTT NAEP" program. Beginning in 1990, however, the NAEP program evolved into two separate assessment programs.

Main NAEP Program

The main NAEP program was first administered in 1990. In 1996, NAGB[14] issued a policy statement to redesign the NAEP.[15] The most noteworthy change was splitting the NAEP into two "unconnected" assessment programs. NAGB proposed a "main NAEP" program that would become the primary way to measure reading, mathematics, science, and writing. NAGB recognized, however, that the nation's curricula would continue to change over time and there would still be value in tracking long-term trends with a stable assessment. NAGB, therefore, proposed the LTT NAEP assessment would be continued, though less frequently, to track trends over time. The main NAEP assessment framework was expected to change about every decade to account for changes in the nation's curricula while the LTT NAEP assessment framework was set to be stable over time.

Another noteworthy change of the 1996 policy statement was the development of performance standards for the main NAEP. Although the original NAEP had numeric performance levels (i.e., 150, 200, 250, 300, 350), there were no descriptive performance standards associated with these levels (i.e., basic, proficient, and advanced). Performance standards were introduced with the administration of the main NAEP in 1990. The standards were subsequently amended several times over the next five years. In 1996, NAGB committed to improving the performance standards and recommended the continued use of performance standards. Because of this policy

[12] The NAEP assesses the content areas of reading; mathematics; science; writing; the arts; civics; economics; geography; U.S. history; and technology and engineering literacy (TEL).

[13] NAEPAA, Section 303.

[14] NAGB was created by Congress in 1988 and is currently authorized under NAEPAA, Section 302. NAGB is responsible for setting policy for NAEP.

[15] *Redesigning the National Assessment of Educational Progress Policy Statement,* adopted August 2, 1996, available at https://www.nagb.gov/content/nagb/assets/documents/policies/Redesigning%20the%20National%20Assessment%20of%20Educational%20Progress.pdf.

shift, the main NAEP and its subprograms continue to use basic, proficient, and advanced as their performance levels.[16]

The main NAEP has evolved over time and split into several subprograms: national, state, and TUDA. The national NAEP assesses the widest range of subject areas. For the national NAEP, a sample is selected from public and private schools and students, creating a representative sample across the nation.

The state NAEP program[17] began as a trial assessment program in 1990 and currently assesses four subject areas: reading, mathematics, writing, and science. In 1996, the state NAEP program was no longer considered a trial and it included 43 states and jurisdictions.[18] In 2001, there was a significant change to the state NAEP program due to the reauthorization of the ESEA by the No Child Left Behind Act (NCLB; P.L. 107-110). The NCLB required states that receive Title I-A funding to participate in biennial NAEP assessments of reading and mathematics in 4th and 8th grades, provided that the Secretary of Education pays for the testing. Although states receiving funding are required to participate, only a sample of schools, and a sample of students within those schools, are selected from each state to participate, creating a representative sample of students within each participating state. Participation is voluntary at the individual level. The assessments administered in the state NAEP program are exactly the same as the national NAEP assessments. The latest reauthorization of the ESEA retained the requirement that states receiving Title I-A funding to participate in these assessments. In 2017, the most recent administration of the main NAEP, 585,000 4th and 8th grade students participated.

The TUDA program assesses four subject areas: reading, mathematics, writing, and science. The TUDA[19] began in 2002 with six participating districts.[20] Participation has grown with each administration, and in 2017, 27 districts voluntarily participated. A total of 66,500 students participated in the 2017 mathematics assessment and 65,300 students participated in the 2017 reading assessment. The assessments administered in the districts are exactly the same as the national and state NAEP assessments.

LTT NAEP Program

Although it was not called the LTT NAEP program at the time, the LTT NAEP program is typically considered to date back the origin of NAEP in 1969. Since it was initially the only NAEP assessment, the LTT NAEP assessment items changed over time throughout the 1970s and 1980s to reflect changes in the nation's curricula. Since 1990, however, the LTT NAEP program

[16] The main NAEP includes the national, state, and TUDA subprograms. NAGB defines NAEP achievement levels as follows: *Basic* denotes partial mastery of the knowledge and skills that are fundamental for proficient work at a given grade. *Proficient* represents solid academic performance for the given grade level and competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter. *Advanced* presumes mastery of both the *Basic* and *Proficient* levels and represents superior academic performance. For more information, see https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_achiev.aspx.

[17] For more information on the state NAEP program, see https://nces.ed.gov/nationsreportcard/about/state.aspx.

[18] For more information on the history of state participation in NAEP, see https://nces.ed.gov/nationsreportcard/about/state.aspx.

[19] For more information on the TUDA program, see https://nces.ed.gov/nationsreportcard/about/district.aspx.

[20] From 1988 to 1994, "below-state" use of NAEP was prohibited. The ESEA, as amended by the Improving America's Schools Act (IASA; P.L. 103-382) removed this prohibition. The current authorizing legislation, the National Assessment of Educational Progress Assessment Act (NAEPAA; P.L. 107-279, Title III) continues to allow "below-state" uses of NAEP.

has remained unchanged. This continuity of assessment items over time is what allows the LTT to accurately track long-term trends. In early administrations of the LTT NAEP program, a wide range of content areas was assessed, including reading, mathematics, science, writing, citizenship, literature, social studies, music, art, and several areas of basic skills. In 1999, due to the development and administration of the main NAEP program, the LTT began to assess only reading and mathematics.[21]

The LTT NAEP program currently assesses 9-, 13-, and 17-year-old students in reading and mathematics. The LTT NAEP was most recently administered in 2012 to approximately 53,000 students. While previously administered about every four years, the next LTT NAEP administration is scheduled for 2024.[22]

U.S. Performance on the National Assessment of Educational Progress

ED provides reports and data tools to explore the results of the NAEP. For example, ED releases publications and multimedia materials for educators, researchers, news organizations, and the public.[23] ED also provides access to the NAEP Data Explorer, which allows the public to create customizable tables and graphics by state, district, content area, etc.[24] The NAEP Data Explorer can also be used to conduct basic research analyses of NAEP data, such as significance testing, gap analysis, and regression analysis.

Due to the amount of information provided by NAEP publications and the NAEP Data Explorer, it is not feasible to cover all NAEP results in this report. This discussion of NAEP results presented here focuses on major trends in performance over time as well as some recent trends. These trends are examined in terms of average scores across groups and average scores across groups of different achievement levels (i.e., 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of achievement). Average trends across different achievement levels are often examined to determine whether the improvement (or lack thereof) can be attributed to higher-achieving students, lower-achieving students, or all students. Additionally, achievement gaps over time that are reported in NAEP publications are also presented. Results discussed herein are used in subsequent sections of this report to highlight some of the issues of interpretation in large-scale assessments. For links to more comprehensive results for NAEP, see **Appendix B**.

Highlights from 2017 Main NAEP

The most recent administration of the main NAEP was 2017. **Figure 1** shows the mathematics and reading results for 4^{th} and 8^{th} graders.

[21] The National Assessment Governing Board (NAGB) officially changed its policy on the LTT NAEP in 2002. For more information, see the *National Assessment Governing Board Long-Term Trends Policy Statement*, adopted May 18, 2002, available at https://www.nagb.gov/content/nagb/assets/documents/policies/Long-term%20Trend.pdf. For more information on the differences between LTT and main NAEP, see https://nces.ed.gov/nationsreportcard/about/ltt_main_diff.aspx. For information on the content areas administered by year from 1969 to 2024, see the NAEP Assessment Schedule, available at https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx.

[22] For a copy of the NAEP schedule of assessments, see https://nces.ed.gov/nationsreportcard/about/assessmentsched.aspx.

[23] For more information, see https://nces.ed.gov/nationsreportcard/pubs_newsroom/.

[24] For more information, see https://nces.ed.gov/nationsreportcard/about/naeptools.aspx.

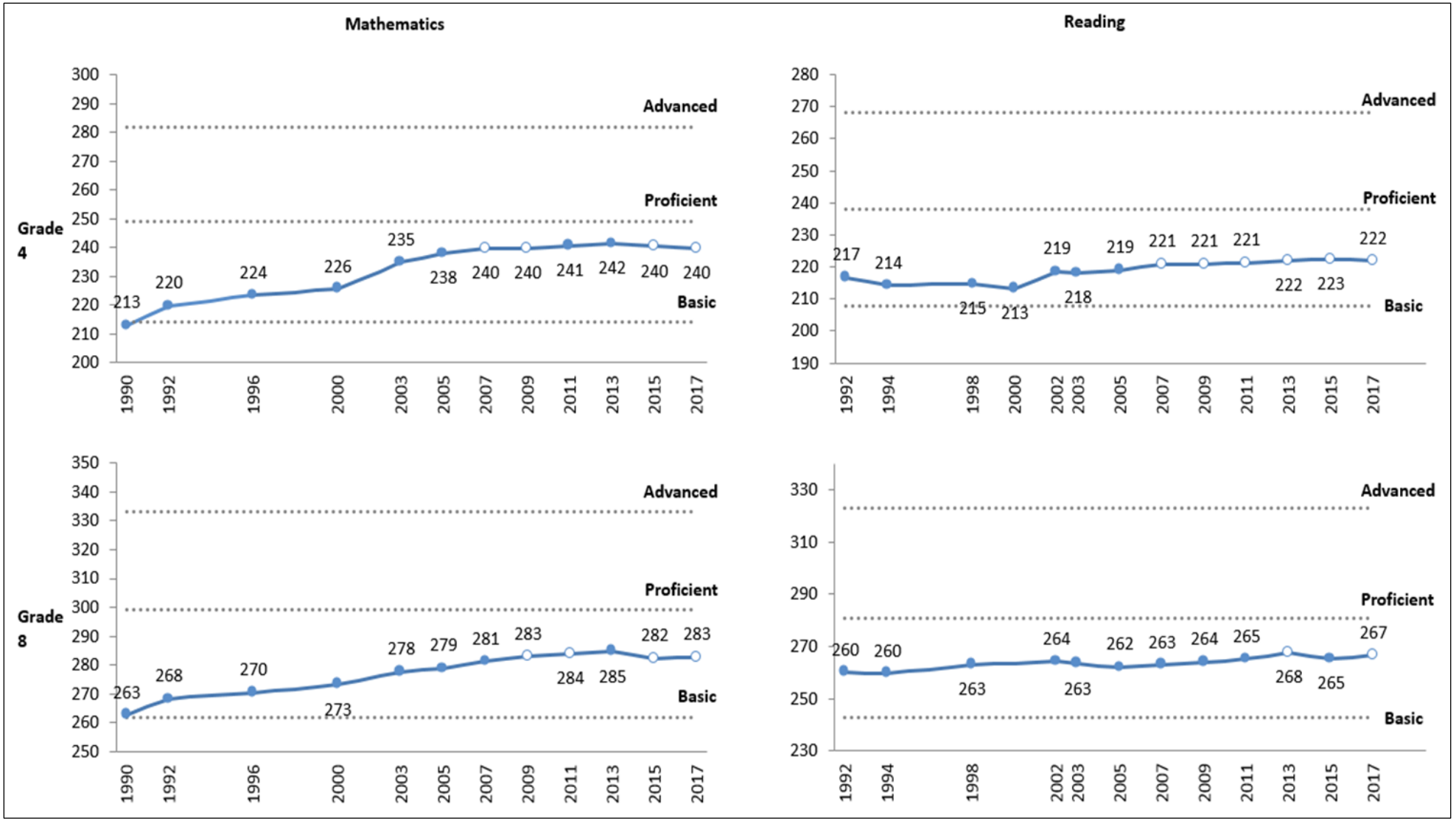
- Average scores have increased significantly in mathematics and reading performance for 4th and 8th grade on the main NAEP (since 1990 and 1992, respectively).
- Compared to the 2015 administration of the NAEP, average *mathematics* scores did not change significantly for 4th or 8th grade. Average *reading* scores did not change significantly for 4th grade students, but there was a small, statistically significant improvement for 8th grade students.[25]

Statistical Significance in Assessment Score Results Reported in Figures

The figures in this section of the report present trend lines of data points. The differences between certain assessment results are tested for statistical significance. All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure.

[25] By convention, results of large-scale assessments are discussed in terms of "statistical significance." While a two-point increase may or may not be of educational significant, it is highlighted because of its statistical significance. The potential misalignment of statistical significance and educational significance is discussed in a later section, "The 'Significance' of Assessment Results."

Figure 1. Average Main NAEP Performance, 1990 to 2017



Source: U.S. Department of Education, *2017 NAEP Mathematics & Reading Assessments: Highlighted Results at Grades 4 and 8 for the Nation, States, and Districts*, National Scores at a Glance, https://www.nationsreportcard.gov/reading_math_2017_highlights/.

Notes: All significance tests are relative to the final year of administration. All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. For more information on NAEP performance standards, see https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_achiev.aspx.

NAEP uses three performance standard levels to describe achievement: basic, proficient, and advanced.[26] For all grades and subject areas, U.S. students' *average* performance in 2017 falls between the basic and proficient levels. For the NAEP assessment, the proficient level of achievement is not considered "grade-level work." The proficient level is considered mastery of challenging subject matter, including the application of knowledge and demonstration of analytical skills.[27]

- For 4th grade *mathematics*, 40% of students scored at or above the proficient level, which is not significantly different than the previous administration in 2015 (40%) but significantly higher than the initial administration in 1990 (13%).[28]
- For 8th grade mathematics, 34% of students scored at or above the proficient level, which is not significantly different than the previous administration in 2015 (33%) but significantly higher than the initial administration in 1990 (15%).[29]
- For 4th grade *reading*, 37% of students scored at or above the proficient level, which is not significantly different than the previous administration in 2015 (36%) but significantly higher than the initial administration in 1992 (29%).[30]
- For 8th grade *reading*, 36% of students scored at or above the proficient level, which is significantly higher than the previous administration in 2015 (34%) and the initial administration in 1992 (29%).[31]

Main NAEP scores are also reported at five different percentiles to track the performance of lower-achieving students, average-achieving students, and higher-achieving students over time (i.e., 10th, 25th, 50th, 75th, and 90th percentiles). **Figure 2** shows the progress over time for students achieving at various percentiles.

- Significant gains on the main NAEP assessment in the last several years are driven by students who are in higher-achieving percentile groups.
- 8th grade students in the 75th and 90th percentile groups made significant gains in mathematics and reading since 2015.
- 8th grade students in the 25th percentile group scored significantly lower in mathematics since 2015.
- 4th grade students in the 25th and 10th percentile groups scored significantly lower in mathematics and reading since 2015.[32]

[26] For definitions of NAEP performance standards, see https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_achiev.aspx.

[27] Some argue that the NAEP proficient level is significantly above grade level. For example, see Tom Loveless, *The NAEP proficiency myth*, Brookings Institution, June 12, 2006, https://www.brookings.edu/blog/brown-center-chalkboard/2016/06/13/the-naep-proficiency-myth/.

[28] https://www.nationsreportcard.gov/math_2017/#/nation/achievement?grade=4.

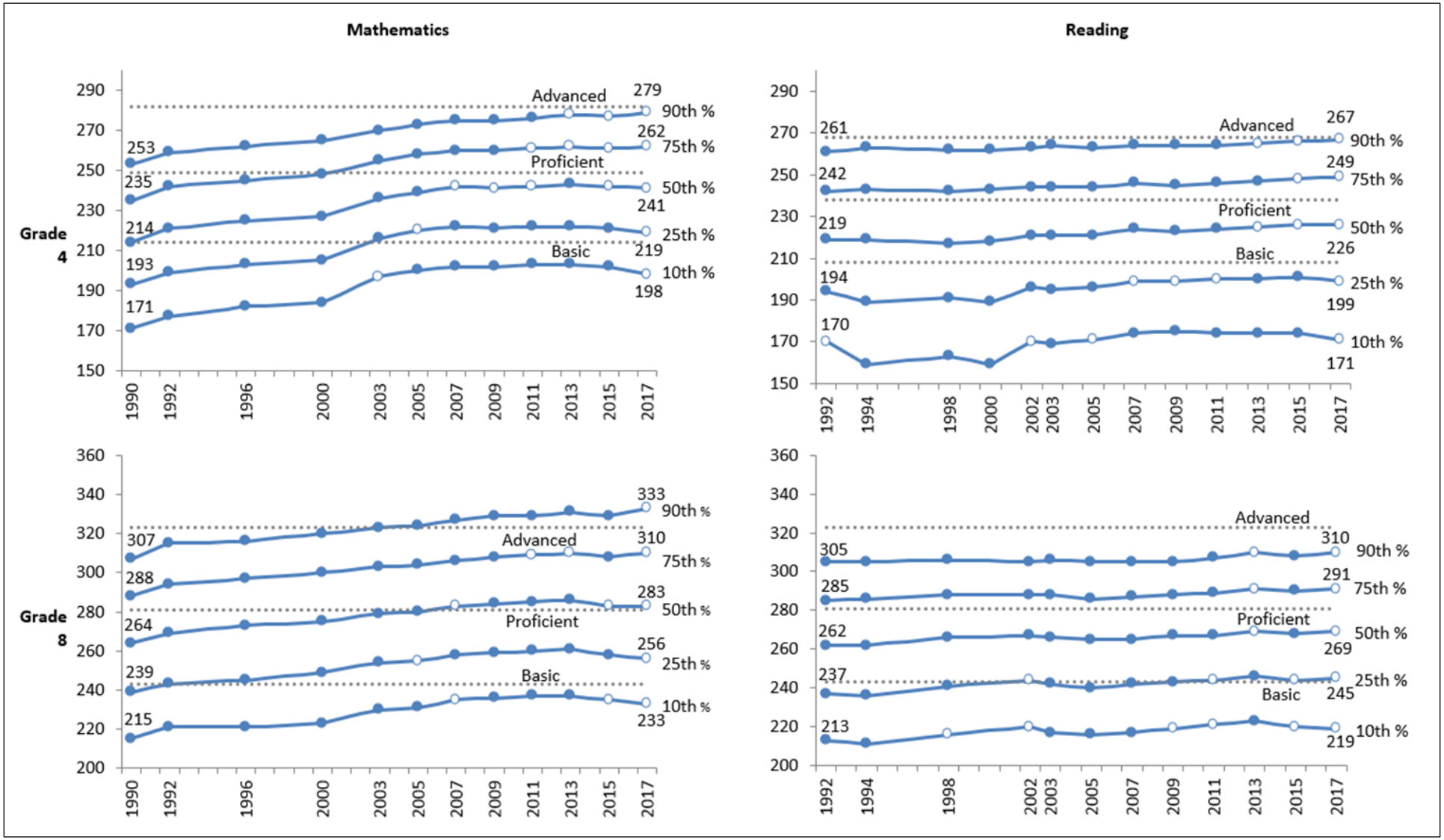
[29] https://www.nationsreportcard.gov/math_2017/#nation/achievement?grade=8.

[30] https://www.nationsreportcard.gov/reading_2017/#/nation/achievement?grade=4.

[31] https://www.nationsreportcard.gov/reading_2017/#nation/achievement?grade=8.

[32] For more information, see https://www.nationsreportcard.gov/reading_math_2017_highlights/.

Figure 2. Average Main NAEP Performance, 1990-2017, by Percentile of Achievement



Source: U.S. Department of Education, *2017 NAEP Mathematics & Reading Assessments: Highlighted Results at Grades 4 and 8 for the Nation, States, and Districts*, National Scores at a Glance, https://www.nationsreportcard.gov/reading_math_2017_highlights/.

Notes: All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. For more information on NAEP performance standards, see https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_achiev.aspx.

Highlights from LTT NAEP

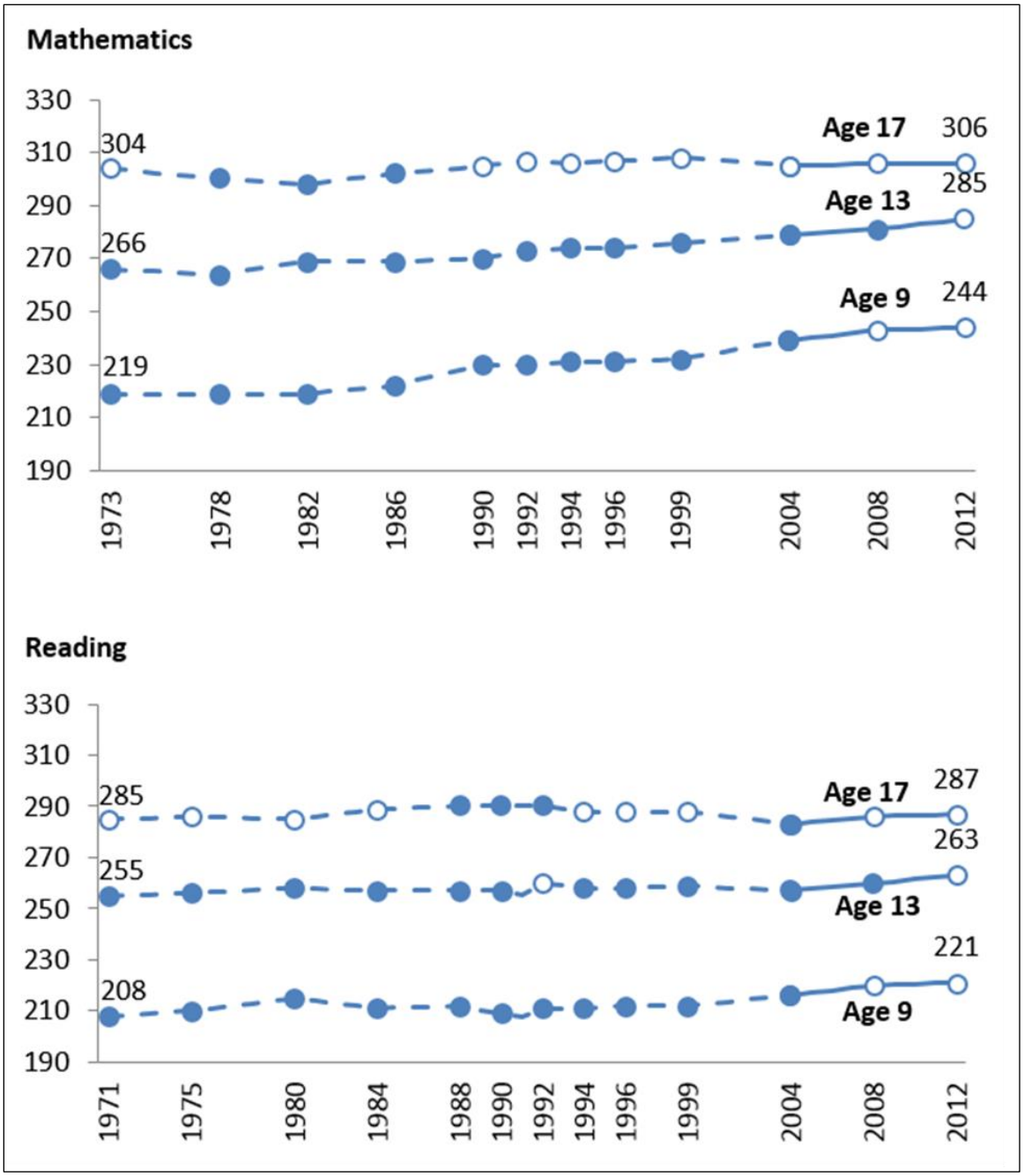
Figure 3 shows the trend in average NAEP mathematics and reading performance on the LTT from the early 1970s until the most recent assessment in 2012.

- The LTT NAEP assessment corroborates the gains observed on the main NAEP for 9- and 13-year-old students.[33]
- In both mathematics and reading, 9- and 13-year-old students have shown significant gains over time.[34]

[33] The next administration of the LTT NAEP is scheduled for 2024. After the results from 2024 are reported, trends across time for different levels of achievement can be tracked across the main NAEP and LTT NAEP within the same timeframe. Results from both assessments may be helpful to identify whether low-achieving groups are not increasing at the same rate as higher-achieving groups. Like the main NAEP, the LTT NAEP scores are also reported at five different percentiles to track the performance of lower-achieving students, average-achieving students, and higher-achieving students over time (i.e., 10th, 25th, 50th, 75th, and 90th percentiles). Unlike the main NAEP, the increases over time on the LTT NAEP do not seem to be driven by higher-achieving groups, but rather across all levels of achievement. See https://www.nationsreportcard.gov/reading_math_2017_highlights/.

[34] The results for 17-year-old students are not compared to main NAEP assessments since comparable data was not presented in this report.

Figure 3. LTT NAEP Average Mathematics and Reading Scores for 9-, 13-, and 17-year-old Students



Source: U.S. Department of Education, *NAEP 2012: Trends in Academic Progress*, NCES 2013-456, 2013, https://nces.ed.gov/nationsreportcard/subject/publications/main2012/pdf/2013456.pdf.

Notes: All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure.

Achievement Gaps Reported by Main NAEP

Achievement gaps occur when one subgroup of students significantly outperforms another subgroup of students on an assessment of academic achievement. In the United States, there have historically been observed achievement gaps by gender, race, ethnicity, socioeconomic status, and disability status. NAEP results have highlighted various achievement gaps and tracked them over time. The following section reports selected achievement gaps that are often highlighted in publications presented by ED. This section does not, however, examine all possible achievement gaps.[35]

The 2017 NAEP results reveal that significant achievement gaps exist by gender, race, ethnicity, socioeconomic status, and school factors. **Table 2** shows the size of the most recent gaps. The largest achievement gaps are typically by race, ethnicity, and socioeconomic status. For the 2017 NAEP results,

- the largest significant gap reported is that between white students and black students in 8th grade mathematics (32 points),
- the second-largest significant gap reported is that between students not eligible for the National Student Lunch Program (NSLP) and students who are eligible for the program in 8th grade mathematics,[36] and
- the smallest significant achievement gaps reported are between male students and female students in mathematics.[37]

Table 2. NAEP Achievement Gaps Across Subgroups, Content Areas, and Grade Levels

	Mathematics		Reading	
2017 Score Gap	4th Grade	8th Grade	4th Grade	8th Grade
Male-Female	2 points	1 point	-6 points	-10 points
White-Black	25 points	32 points	26 points	25 points
White-Hispanic	19 points	24 points	23 points	19 points
Asian/Pacific Islander-White	10 points	17 points	7 points	7 points
Not Eligible for NSLP-Eligible for NSLP	24 points	29 points	29 points	24 points
Catholic-Public	6 points	12 points	14 points	18 points
Other Noncharter Public Schools-Charter Schools	4 points	1 point[a]	No difference[b]	No difference[b]

[35] The NAEP Data Explorer can be used to examine other achievement gaps of interest. See https://nces.ed.gov/nationsreportcard/data/.

[36] Eligibility for the National Student Lunch Program is often used as a proxy measure for the percentage of students living in poverty in educational research. For more information, see U.S. Department of Education, National Center for Education Statistics, NCES Blog, *Free or reduced price lunch: A proxy for poverty?*, April 16, 2015, https://nces.ed.gov/blogs/nces/post/free-or-reduced-price-lunch-a-proxy-for-poverty.

[37] In 4th grade, male students outperform female students by 2 points and in 8th grade males outperform females by 1 point. While these achievement gaps are statistically significant, they are considerably smaller in magnitude than achievement gaps by ethnicity or socioeconomic status.

Source: U.S. Department of Education, *2017 NAEP Mathematics & Reading Assessments: Highlighted Results at Grades 4 and 8 for the Nation, States, and Districts*, National Scores at a Glance, https://www.nationsreportcard.gov/reading_math_2017_highlights/.

Notes: All achievement gaps included in the table are statistically significant, unless otherwise noted.

a. No significant difference.

b. Rounds to zero.

Some achievement gaps have changed since the early 1990s. As reported by ED, some have increased significantly over time.

- The largest increase in the achievement gap over time is the difference between white students and Asian/Pacific Islander students in 4th grade reading (15 points) and 8th grade mathematics (12 points). In 4th grade reading, white students outperformed Asian/Pacific Islander students in 1992 but are now significantly outperformed by them. In 8th grade mathematics, Asian/Pacific Islander students outperformed white students in 1990 and the gap has become significantly larger over time.[38]

Other achievement gaps have decreased significantly over time.

- The gap between white students and black students has decreased in both 4th grade mathematics (7 points) and 4th grade reading (6 points).
- The gap between white students and Hispanic students has decreased in 8th grade reading (7 points).[39]

Achievement Gaps Reported by LTT NAEP

The LTT NAEP also tracks achievement gaps over time. In general, achievement gaps have significantly narrowed or remained unchanged. For example, the gap between white students and black students in reading at age 9 has narrowed since 1971. While the average score for white students increased 15 points, the average score for black students increased 36 points, leading to the narrowing of the achievement gap.[40] None of the measured achievement gaps in the LTT NAEP have increased significantly over time.

International Assessments

The United States regularly participates in three international assessments: TIMSS, PIRLS, and PISA. While U.S. students have participated in international assessments since the 1960s, the modern era of international assessments began in the mid-1990s.[41] This report focuses on international assessment results that highlight U.S. student performance over time and in relation to other countries. Results discussed herein are used in subsequent sections of this report to

[38] For more information, see https://www.nationsreportcard.gov/reading_math_2017_highlights/.

[39] For more information, see https://www.nationsreportcard.gov/reading_math_2017_highlights/.

[40] See https://nces.ed.gov/nationsreportcard/pubs/main2012/2013456.aspx#section2.

[41] For example, the United States participated in the First International Mathematics Study (FIMS) in 1964, the Second International Mathematics Study (SIMS) in 1981-1982, and the Third International Mathematics and Science Study (TIMSS) in 1994-1995. As TIMSS continued to be administered, the name was changed into the Trends in International Mathematics and Science Study (also TIMSS). The current TIMSS began in 1995 and is administered every four years.

highlight some of the issues of interpretation in large-scale assessments. For links to more comprehensive results for the international assessments, see **Appendix B**.

TIMSS

The TIMSS is an international comparative study that is designed to measure mathematics and science achievement in 4th and 8th grades. The TIMSS is designed to measure "school-based learning," and is designed to be broadly aligned with mathematics and science curricula in participating education systems (i.e., countries and some subnational jurisdictions).[42] The United States has participated in the TIMSS every four years since 1995. Less often, 12th grade students participate in the TIMSS Advanced program, which measures advanced mathematics and physics.[43] In 2015, approximately 20,250 U.S. students participated in TIMSS and about 5,900 U.S. students participated in TIMSS Advanced. The United States was one of over 60 education systems to participate in TIMSS and one of 9 to participate in the TIMSS Advanced program.[44] All participation in TIMSS is voluntary. The next TIMSS administration is scheduled for 2019. No date has been announced for the next TIMSS Advanced administration.

The TIMSS is conducted in the United States under the authority of international assessment activities.[45] TIMSS assessments in the United States are administered by the Commissioner of NCES within the International Activities Program. The International Association for the Evaluation of Educational Achievement (IEA) coordinates TIMSS and TIMSS Advanced internationally.

U.S. Performance on TIMSS in Relation to Other Countries[46]

TIMSS reports results separately for mathematics and science. U.S. results for TIMSS mathematics are as follows:[47]

- In 2015, 4th grade, the United States scored significantly lower than 10 education systems, significantly higher than 34 education systems, and not significantly different than 9 education systems.[48]

[42] TIMSS also includes a "TIMSS Advanced" assessment, which assesses advanced mathematics and physics achievement of students in their final year of high school. The United States has participated in TIMSS Advanced two times (1995 and 2015). Compared to TIMSS, the TIMSS Advanced assessment program has few participating countries and data do not yet allow for long-term trends of educational progress to be observed. Results of TIMSS Advanced are outside of the scope of this report. For more information on TIMSS Advanced, see U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[43] The TIMSS Advanced also focuses school-based learning rather than real-world application of skills.

[44] For a list of participating countries, see U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[45] ESRA, Section 153(a)(6).

[46] Results discussed herein focus on TIMSS and do not describe the results of TIMSS Advanced. Since TIMSS Advanced is a smaller assessment with fewer participating countries and a shorter history of administration, it may be more informative to focus on TIMSS. For more information on the results of TIMSS Advanced, see U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[47] In the discussion of international assessment results, "not measurably different" means that there is no statistically significant difference between the average score of U.S. students and the average score of other education systems.

[48] The United States average score ranks #15; however, four of the education systems scoring above the United States

- In 8th grade, the United States scored significantly lower than 8 education systems, significantly higher than 24 education systems, and not significantly different than 10 education systems.[49]

U.S. results for TIMSS science are as follows:

- In 2015, 4th grade, the United States scored significantly lower than 7 education systems, significantly higher than 38 education systems, and not significantly different than 7 education systems.[50]
- In 8th grade, the United States scored significantly lower than 7 education systems, significantly higher than 26 education systems, and not significantly different than 9 education systems.[51]

U.S. Performance on TIMSS over Time by Achievement Percentiles

TIMSS reports results over time by achievement level for U.S. students (i.e., 10th percentile, 25th percentile, 50th percentile, 75th percentile, and 90th percentile). **Figure 4** shows the results for 4th and 8th grade mathematics for U.S. students.

- Increases in achievement for 4th grade mathematics may be driven by the performance of average or above average groups, however, the increases are not statistically significant.[52]
- Performance on TIMSS increased for all achievement levels on 8th grade mathematics, however, the increases were significant for average and above average groups while increases were not significant for below average groups.

had average scores that were not statistically significantly higher and five education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Figure 1a. Average mathematics scores of 4th-grade students, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

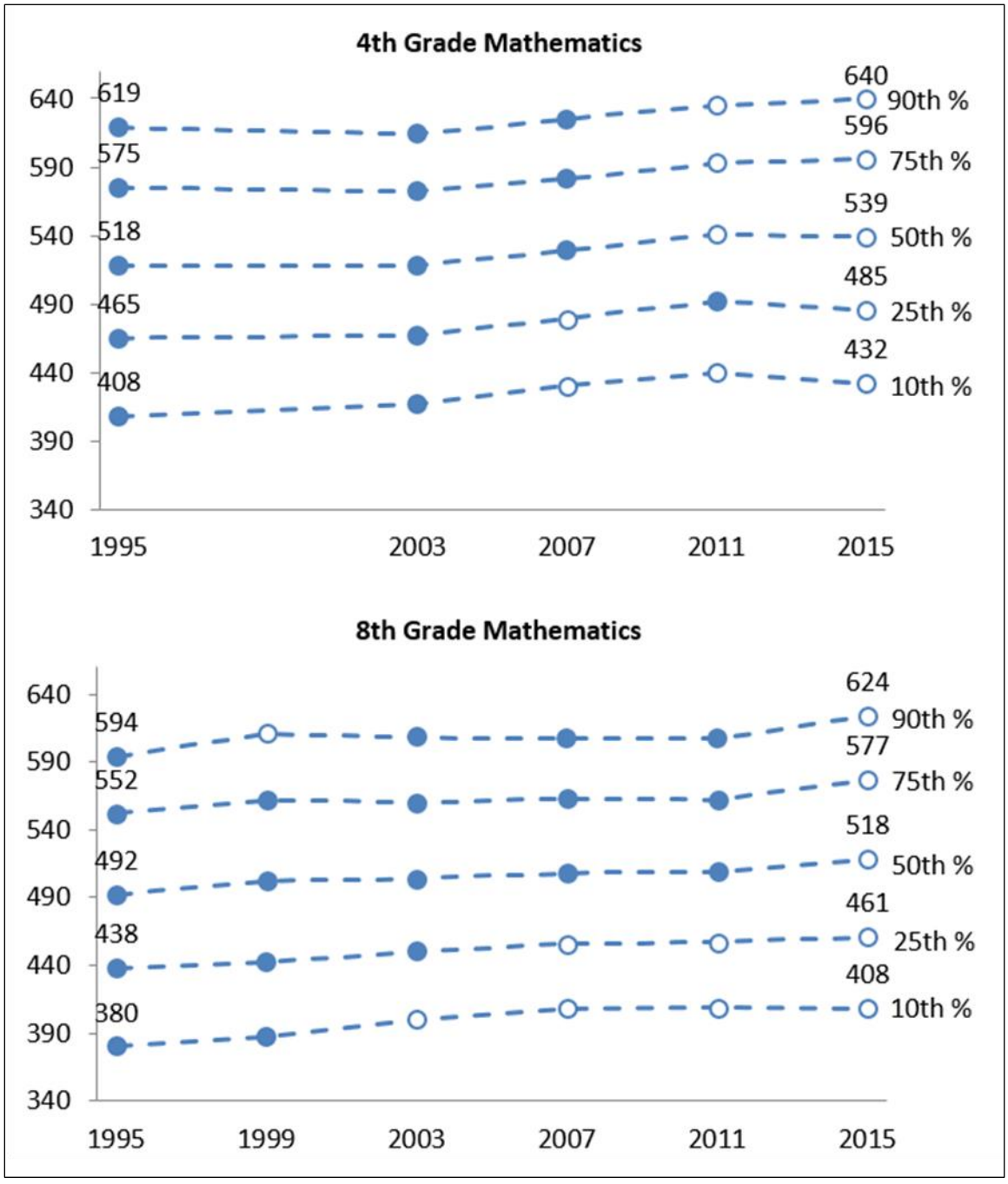
[49] The United States average score ranks #12; however, three of the education systems scoring above the United States had average scores that were not statistically significantly higher and seven education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Figure 1b. Average mathematics scores of 8th grade students, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[50] The United States average score ranks #11; however, three of the education systems scoring above the United States had average scores that were not statistically significantly higher and four education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education systems rankings, see "Figure 5a. Average science scores of 4th grade students, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[51] The United States average score ranks #11; however, three of the education systems scoring above the United States had average scores that were not statistically significantly higher and six education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Figure 5b. Average science scores of 8th grade students, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

[52] This trend is similar to the NAEP results trend between 2015 and 2017.

Figure 4. Trends in U.S. 4th and 8th Grade TIMSS Average Mathematics Scores by Achievement Level and Year



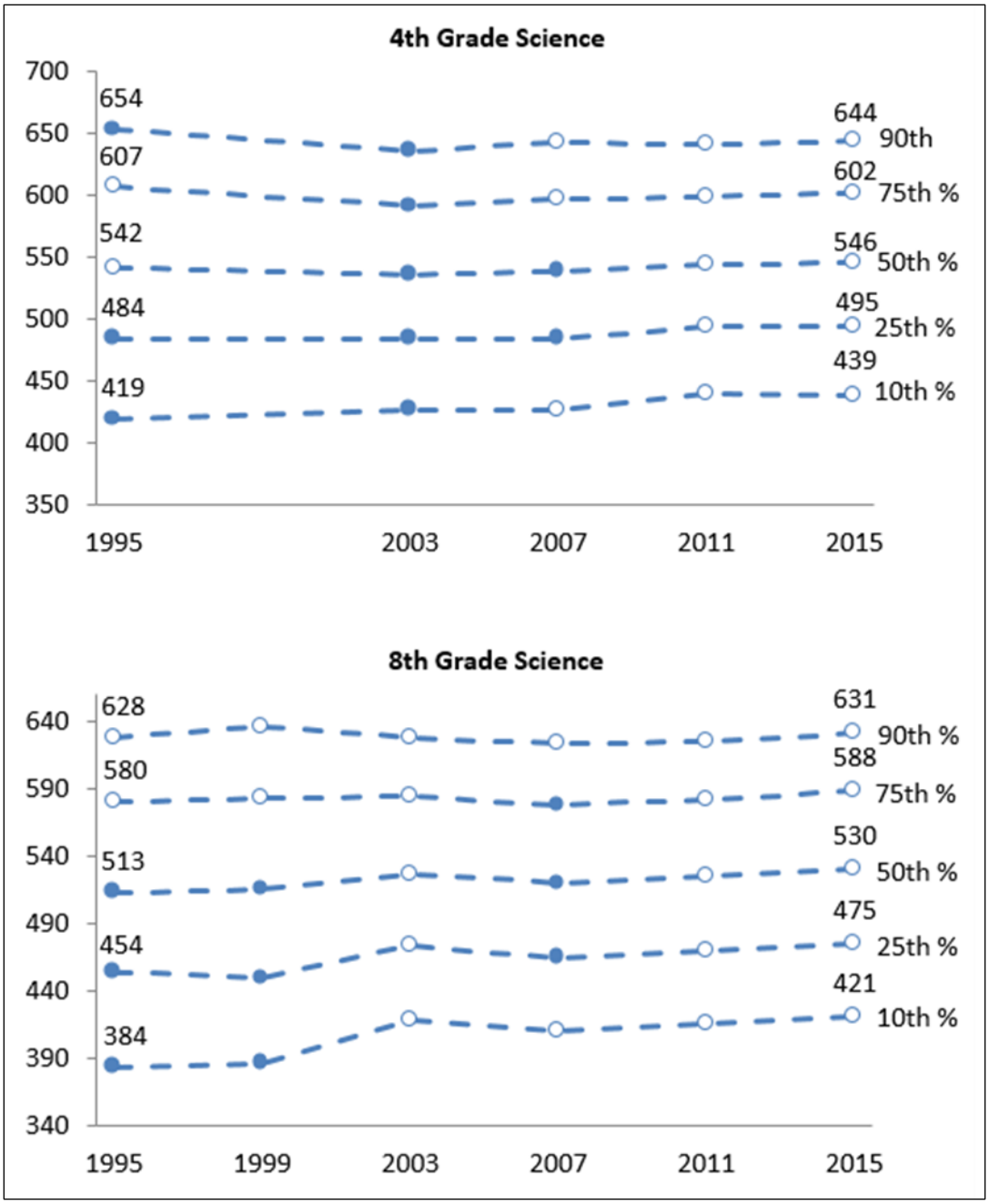
Source: U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

Notes: All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure.

Science results for U.S. students are also reported over time and by achievement level. **Figure 5** shows results for science performance for students in 4th and 8th grade over time and by achievement level.

- Science achievement in 4th and 8th grades has been generally flat since the 2011 administration of TIMSS.
- There have been some significant increases in 4th and 8th grade science achievement since the 2007 administration of TIMSS for students whose achievement falls between the 25th and 75th percentiles.

Figure 5. Trends in U.S. 4th and 8th Grade TIMSS Average Science Scores by Achievement Level and Year



Source: U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

Notes: All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure.

PIRLS

PIRLS is an international comparative study of 4th grade students in reading literacy. PIRLS assesses reading literacy at 4th grade because this is typically considered a developmental stage of learning where students shift from learning to read to reading to learn. PIRLS is not an assessment of word reading ability but rather an assessment of the purposes for reading, processes of comprehension, and reading behavior and attitudes. For young students, reading generally has two purposes both in and out of school: (1) reading for literacy experience, and (2) reading to acquire and use information.

The United States has participated in PIRLS every five years since 2001. The next assessment of PIRLS will be administered in 2021. In 2016, the United States also participated in the first administration of ePIRLS, a computer-based assessment of online reading. ePIRLS is designed to measure informational reading comprehension skills in an online environment.[53] In 2016, approximately 4,500 U.S. students participated in PIRLS and an additional 4,000 students participated in ePIRLS. The United States was one of 61 education systems to participate in PIRLS and one of 14 to participate in ePIRLS. All participation in PIRLS and ePIRLS is voluntary.

PIRLS is conducted in the United States under the authority of international assessment activities.[54] The PIRLS and ePIRLS assessments in the United States are administered by the Commissioner of NCES within the International Activities Program. Like TIMSS, the IEA coordinates PIRLS internationally.

U.S. Performance on PIRLS and ePIRLS in Relation to Other Countries

Results for PIRLS and ePIRLS are reported separately.[55]

- For PIRLS, the United States scored significantly lower than 12 education systems, significantly higher than 30 education systems, and not significantly different than 15 education systems.[56]
- For ePIRLS, the United States scored significantly lower than 3 education systems, significantly higher than 10 education systems, and not significantly different than 2 education systems.[57]

[53] The next administration of ePIRLS has not been announced.

[54] ESRA, Section 153(a)(6).

[55] In the discussion of international assessment results, "not measurably different" means that there is no statistically significant difference between the average score of U.S. students and the average score of other education systems.

[56] The United States average score ranks #15; however, three of the education systems scoring above the United States had average scores that were not statistically significantly higher, eleven education systems scoring below the United States had average scores that were not statistically significantly lower, and one education system had the same average score. To see the education system rankings, see "Table 1. PIRLS overall reading average scale scores of fourth-grade students, by education system: 2016" in U.S. Department of Education, National Center for Education Statistics, *Reading Achievement of U.S. Fourth-Grade Students in an International Context*, NCES 2018-017, December 2017, https://nces.ed.gov/pubs2018/2018017.pdf.

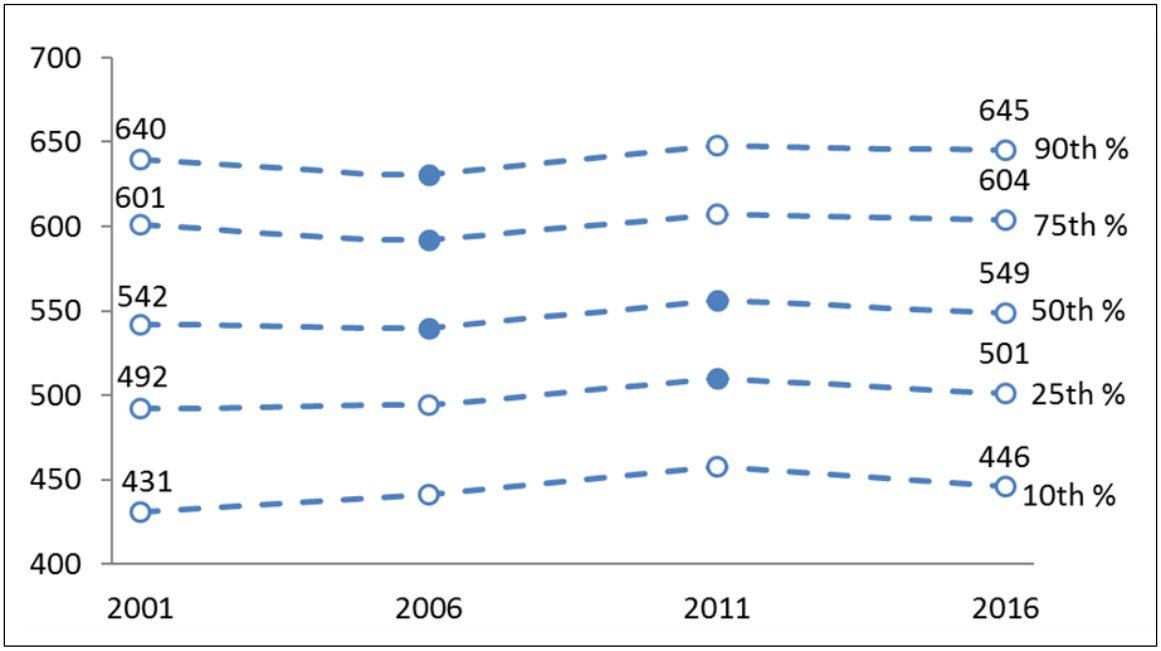
[57] The United States average score ranks #6; however, two of the education systems scoring above the United States had average scores that were not statistically significantly higher. To see the education system rankings, see "Table 3. ePIRLS online informational reading average scale scores of fourth-grade students, by education system: 2016" in U.S. Department of Education, National Center for Education Statistics, *Reading Achievement of U.S. Fourth-Grade Students in an International Context*, NCES 2018-017, December 2017, https://nces.ed.gov/pubs2018/2018017.pdf.

U.S. Performance on PIRLS over Time by Achievement Levels

PIRLS reports results for U.S. students over time by achievement level (i.e., 10th percentile, 25th percentile, 50th percentile, 75th percentile, and 90th percentile).[58] **Figure 6** shows the results for 4th grade reading achievement by achievement level across time.

- In general, U.S. student performance from 2001 to 2016 was relatively flat.

Figure 6. Trends in U.S. 4th Grade Average PIRLS Reading Scores by Achievement Level and Year



Source: U.S. Department of Education, National Center for Education Statistics, *Reading Achievement of U.S. Fourth-Grade Students in an International Context*, NCES 2018-017, December 2017, https://nces.ed.gov/pubs2018/2018017.pdf.

Notes: All significance tests are relative to the last year of assessment administration included in the figure and represented by an open data point. Any *solid data point* along the trend line indicates a *statistically significant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure. Any *open data point* along the trend line indicates a *statistically insignificant difference* between that year's assessment results and the assessment results for the last year of assessment administration included in the figure.

PISA

PISA is an international comparative study of 15-year-old students in the content areas of science, reading, and mathematics "literacy." It aims to measure the achievement of students at the end of their compulsory education.[59] The PISA is not designed to measure "school-based learning" and is not designed to be aligned with academic content standards. Instead, PISA intends to measure students' preparation for life and focuses on science, reading, and mathematics problems within a

[58] Since 2016 was the first administration of ePIRLS, there are no results available to examine trends over time.

[59] The end of compulsory education in the United States is typically older than age 15 and is determined by state. To see minimum and maximum ages for compulsory education by state, see https://nces.ed.gov/programs/statereform/tab5_1.asp.

real-life context.[60] The United States has participated in PISA every three years since 2000. In 2015, approximately 6,000 U.S. students participated in PISA. The United States was one of 72 countries and economies to participate. All participation is voluntary. PISA 2018 was administered in the fall of 2018, and results are tentatively scheduled to be released in December 2019.[61]

PISA is conducted under the authority of international assessment activities.[62] The PISA assessment in the United States is administered by the Commissioner of NCES within the International Activities Program. Unlike TIMSS and PIRLS, the international coordination of the PISA is conducted by the Organisation for Economic Co-operation and Development (OECD), an intergovernmental organization of industrialized countries.

U.S. Performance on PISA in Relation to Other Countries

PISA reports results separately for reading literacy, mathematics literacy, and science literacy.[63]

- For reading literacy, in 2015, the United States scored significantly lower than 14 education systems, significantly higher than 42 education systems, and not significantly different than 13 education systems.[64]
- For mathematics literacy, in 2015, the United States scored significantly lower than 36 education systems, significantly higher than 28 education systems, and not significantly different than 5 education systems.[65]
- For science literacy, in 2015, the United States scored significantly lower than 18 education systems, significantly higher than 39 education systems, and not significantly different than 12 education systems.[66]

---

[60] U.S. Department of Education, National Center for Education Statistics, Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context: First Look at PISA 2015, December 2016, p. 1, https://nces.ed.gov/pubs2017/2017048.pdf.

[61] https://nces.ed.gov/surveys/pisa/schedule.asp.

[62] ESRA, Section 153(a)(6).

[63] [63] In the discussion of international assessment results, "not measurably different" means that there is no statistically significant difference between the average score of U.S. students and the average score of other education systems.

[64] The United States average score ranks #25; however, nine of the education systems scoring above the United States had average scores that were not statistically significantly higher and four education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Table 2. Average scores of 15-year-old students on the PISA reading literacy scale, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context: First Look at PISA 2015*, December 2016, p. 1, https://nces.ed.gov/pubs2017/2017048.pdf.

[65] The United States average score ranks #41; however, three of the education systems scoring above the United States had average scores that were not statistically significantly higher and two education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Table 3. Average scores of 15-year-old students on the PISA mathematics literacy scale, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context: First Look at PISA 2015*, December 2016, p. 1, https://nces.ed.gov/pubs2017/2017048.pdf.

[66] The United States average score ranks #26; however, six of the education systems scoring above the United States had average scores that were not statistically significantly higher and six education systems scoring below the United States had average scores that were not statistically significantly lower. To see the education system rankings, see "Table 1. Average scores of 15-year-old students on the PISA science literacy scale, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context: First Look at PISA 2015*, December 2016, p. 1, https://nces.ed.gov/pubs2017/2017048.pdf.

U.S. Performance on PISA Over Time

Unlike the NAEP and other international assessments, PISA does not track progress over time in the same way for different levels of achievement (e.g., 10th, 25th, 50th, 75th, and 90th percentiles). PISA does, however, track average performance over time. **Table 3** shows average score changes for U.S. students in mathematics, reading, and science literacy:

- For mathematics literacy, the average score in 2015 was 11 points lower than the average score in 2012 and 17 points lower than the average score in 2009; however, the average score in 2015 was not measurably different than the average scores in 2003 and 2006.
- For reading literacy, the average score in 2015 was not measurably different than in previous years.
- For science literacy, the average score in 2015 were not measurably different than in previous years.

Table 3. Trends in Average U.S. PISA Scores by Year

Subject	Average Score					Change in Average Score			
	2003	2006	2009	2012	2015	2015-2003	2015-2006	2015-2009	2015-2012
Mathematics	483	474	487	481	470	No change	No change	Decrease[a]	Decrease[a]
Reading	N/A	N/A	500	498	497	N/A	N/A	No change	No change
Science	N/A	489	502	497	496	N/A	No change	No change	No change

Source: U.S. Department of Education, National Center for Education Statistics, *Performance of U.S. 15-Year-Old Students in Science, Reading, and Mathematics Literacy in an International Context*, NCES 2017-048, December 2016, https://nces.ed.gov/pubs2017/2017048.pdf.

a. This was a statistically significant change.

Issues of Interpretation in Large-Scale Assessments

Results of national and international assessments are difficult to interpret for a number of reasons. Perhaps the most difficult issue in the interpretation of large-scale assessments is processing the large volume of data presented in reports. The results provided in the previous section are a small fraction of what is available. These specific results were reported to provide a broad overview of the achievement of U.S. students across a wide range of assessments over time.

When large numbers of results are reported in national and international assessments, it can be challenging to compile assessment results across assessments to determine how well U.S. students are achieving over time and relative to other countries. The purpose of this section of the report is to present a few issues to consider when interpreting national and international assessments. This discussion is not intended to provide a comprehensive list of possible considerations; however, the key issues presented below are pervasive across large-scale assessments.

The "Significance" of Assessment Results

The concept of statistical significance is central to reporting assessment results. When states or countries are presented in a rank order, it is important to note whether differences in rank are statistically significant. For example, as reported in the TIMSS results above, in 4th grade mathematics, the United States scored lower than 10 education systems, higher than 34 education systems, and not measurably different than 9 education systems. When average scores are presented in a rank order, however, the United States is ranked number 15.[67] Four education systems above the United States and five education systems below the United States had average scores that were not statistically significantly different from the United States. Strictly ranking average scores does not account for statistically insignificant differences between average scores.

Statistical significance is an important measure of whether a change is likely to be due to chance. Statistical significance, however, may not be the most important indicator of meaningful change. A statistically significant change in assessment score is a change that is unlikely to be due to chance. Statistical significance, however, is influenced by many factors. For the purposes of this discussion, the most relevant factor that influences statistical significance is sample size. The larger the sample size, the more likely a small change in assessment score will be statistically significant. Recall that national and international assessments sample tens of thousands or hundreds of thousands of U.S. students. Due to large samples, small increases or decreases in academic achievement may be statistically significant. For example, in the most recent NAEP administration, a two-point increase in 8th grade reading performance was statistically significant.

Statistical significance is not the same as educational significance. Educational significance is subjective and dependent on the educational context of the results. Statistical significance cannot determine the magnitude of the difference and whether or not it is of educational significance. For example, as reported above in the NAEP results, there is a statistically significant gap of *1 point* between male and female students in 8th grade mathematics. There is also a statistically significant gap of *32 points* between white and black students in 8th grade mathematics. While

[67] See "Figure 1a. Average mathematics scores of 4th-grade students, by education system: 2015" in U.S. Department of Education, National Center for Education Statistics, *Highlights From TIMSS and TIMSS Advanced 2015*, NCES 2017-002, November 2016, https://nces.ed.gov/pubs2017/2017002.pdf.

both gaps are statistically significant, the gap between white and black students may have more educational significance.

Some researchers argue that statistical significance can be misleading for policy purposes because a statistically significant result may be too small to warrant a change in practices or policies.[68] Educational significance is more subjective and difficult to define when considering assessment results. One way educational researchers have tried to define educational significance or practical significance is by using an effect size.[69] An effect size can better determine the magnitude of an effect, however, there is still no consensus on the magnitude of an effect size that is meaningful in all contexts.[70]

The Narrow Focus on One Assessment

When new national and international assessment results are released, there is a tendency to focus on a single assessment or a single result. A narrow focus on one assessment at one point in time, however, may not provide appropriate context for interpreting the results. Examining differences in results across assessments and trends over time can provide a more meaningful context for interpretation.

For example, PISA results show a statistically significant decrease in mathematics literacy scores from 2012 to 2015. When considered in isolation, this result may indicate that the mathematics achievement of 15-year-old U.S. students is declining. Consider, however, that 8th grade U.S. students made statistically significant gains in mathematics on TIMSS from 2011 to 2015 and showed no change in mathematics on the NAEP from 2015 to 2017. While conflicting results like these can be frustrating, it is important to consider them together as a body of evidence instead of isolated data points. There may be valid reasons that U.S. students' performance decreased on PISA and increased on TIMSS. For example, as discussed above, TIMSS measures more "school-based learning," and U.S. students have historically scored relatively higher on this assessment. Perhaps the content standards and curriculum in place in the United States are more aligned with content assessed by the TIMSS and less aligned with the content assessed by the PISA.

Another issue to consider is the trend over time. For example, although U.S. students' performance on PISA significantly decreased from 2009 to 2015, the 2015 score is not measurably different than the average scores in 2003 or 2006.[71] While any significant decrease

[68] See, for example, Carnoy, M. & Rothestein, R. (2013). *What do International Tests Really Show About U.S. Students Performance?* Economic Policy Institute., available at https://www.epi.org/publication/us-student-performance-testing/ .

[69] An effect size is calculated by finding the difference in average scores between two groups and dividing by the pooled standard deviation. It is a measure of the size of change in standard deviation units. There are different conventions for what makes an effect size small or large based on the type of research being conducted (e.g., intervention research, policy research, etc.). For more information, see Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available at http://ies.ed.gov/ncser/.

[70] Some researchers have published guidance on presenting statistical findings in a way that make the magnitude of the effect more clear. See Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D. (2012). Translating the Statistical Representation of the Effects of Education Interventions into More Readily Interpretable Forms. (NCSER 2013-3000). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available at http://ies.ed.gov/ncser/.

[71] See **Table 3**.

may be cause for concern, it is important to recognize that scores have not decreased significantly since the initial administration of the PISA. While a statistically insignificant change in achievement across 10 to 15 years may not be considered a positive result, the long-term trend presents a different picture of achievement than the short-term trend. Examining trends allows researchers and policymakers to identify policies and practices that were implemented at a certain time that may have contributed to an observed trend.

In general, reports from a single assessment that claim U.S. student achievement has stagnated, increased, or decreased must be interpreted with caution. When one result is reported in isolation, it is easy to make oversimplified conclusions that do not necessarily generalize across assessments and over time.[72]

Socioeconomic Considerations Across Countries

International assessment results are based on a representative sample of students. There are considerable differences in the characteristics of students within certain countries, however, and an accurate representative sample would also reflect these differences. Some of the differences in populations across countries may have considerable implications in the interpretation of assessment results. For example, one difference that has been found to have implications for the interpretation of assessment results is the range of socioeconomic inequality. The United States has a broader income distribution than many of the countries that participate in international assessments. The sample from the United States, therefore, likely has a larger number of students from lower-income families than samples from countries with more concentrated income distributions.

Some researchers argue that "social class inequality," which is largely determined by income, is a major factor in the interpretation of international assessment results.[73] These researchers found that students from lower-income families perform worse than students from higher-income families in every country in their analysis. Since there are more lower-income families in the United States than in the some of the countries it is routinely compared to, researchers argue that the relative performance of U.S. students is actually better than it appears when simply comparing countries' national averages.

In an analysis of 2009 PISA results, researchers found that if U.S. students had an income distribution similar to that of other countries in the analysis, the average reading scores would be higher than those of the other countries and the average math scores would be about the same.[74] Furthermore, these researchers suggest that examining trends for students at varying income distributions over time would be more useful than examining average scores over time.

[72] Several education researchers have written commentary regarding the dangers of "cherry picking" the data on national and international assessments. See, for example, "Betsy DeVos is Half-Right on Test Scores, But Test Scores Alone Don't Make the Case for School Choice" by Grover J. "Russ" Whitehurst, March 12, 2018, available at https://www.brookings.edu/blog/up-front/2018/03/12/betsy-devos-is-half-right-on-test-scores-but-test-scores-alone-dont-make-the-case-for-school-choice/; and "What You Need to Know About the International Test Scores" by Diane Ravitch, December 3, 2013, available at https://www.huffingtonpost.com/diane-ravitch/international-test-scores_b_4379533.html.

[73] M. Carnoy and R. Rothstein, "What Do International Tests Really Show About U.S. Student Performance?" Economic Policy Institute, 2013, https://www.epi.org/publication/us-student-performance-testing/.

[74] Ibid. Researchers analyzed comparative assessment data for three post-industrialized countries similar to the United States: France, Germany, and the United Kingdom.

Comparing Results Across Assessments

Since U.S. students participate in national assessments and several international assessments, there is a natural inclination to want to compare results from one assessment to another, especially when results are released within a short timeframe. The most frequently administered assessments for U.S. students are annual statewide assessments and biennial NAEP assessments. It often appears as if there is overlap in the content, timing, and grade levels assessed, so it begs the question: can NAEP be compared to the results of statewide assessment systems required by the ESEA?

Although U.S. students participate in international assessments less frequently, there is also apparent overlap in the content, timing, and grade levels assessed. This leads to questions such as the following: Can NAEP be compared to international assessments? If NAEP and TIMSS both measure 8th grade mathematics, are those results comparable? If NAEP and PIRLS both measure 4th grade reading, are those results comparable?

The answers to these questions largely depend on the alignment between assessments and the purpose of the comparison. The following section of the report discusses some of the alignment studies that have been conducted and the usefulness of making comparisons across large-scale assessments.

NAEP and Statewide Assessments Comparisons

Both NAEP and statewide assessments measure 4th and 8th grade achievement in reading and mathematics. They both report scaled scores and performance levels of students in these content areas. These similarities may lead some to question whether these assessments are comparable or even redundant. While national and state assessments may appear to have significant similarities, each was designed for a different purpose and by different stakeholders. There are three main issues to contemplate when considering making a comparison: the alignment of content standards, the scale, and the definition of performance standards.

NAEP and statewide assessments overlap in the sense that both assessment programs measure mathematics and reading achievement. The assessment programs, however, use different frameworks to decide what mathematics and reading content will be measured. For NAEP, the NAGB determines what students know and should be able to do in various content areas based on the knowledge and experience of various stakeholders, such as content area experts, school administrators, policymakers, teachers, and parents. The content assessed by NAEP is not aligned to any particular content standards.[75] The specific content measured by statewide assessments, however, is aligned with the state's content standards. Each state has a different process for determining its content standards for mathematics and reading, but, like NAEP, it also includes input from multiple stakeholders.

While it may not be feasible to study the content alignment between NAEP and all states' content standards, there was a recent alignment study between NAEP and the common core state standards (CCSS). The study used an expert panel to study the alignment of NAEP and CCSS in 4th and 8th grade mathematics. The study found 79% alignment for 4th grade students and 87% alignment for 8th grade students, concluding that alignment between NAEP and CCSS was

[75] For more information, see U.S. Department of Education, National Center for Education Statistics, *Comparing NAEP and State Assessments*, https://nces.ed.gov/nationsreportcard/about/comparing_assessments.aspx.

"strong."[76] Other investigations have examined the alignment of NAEP reading and writing frameworks and the CCSS English language arts standards, however, these examinations did not determine a degree of alignment.[77] It is important to note that many states are not currently using the CCSS or are using a modified version of the CCSS. From the data presented here, it is not possible to determine how well the NAEP framework aligns with specific state content standards in mathematics and reading.

Even if a "strong" alignment between NAEP frameworks and state content standards is assumed, there are other difficult issues to consider when making comparisons between the assessment results. For example, NAEP and statewide assessments use different scales. NAEP scaled scores for reading and mathematics are reported on a scale from 0 to 500. Statewide assessments use a scale that is specific to the assessment used in each state. For the purpose of illustration, consider three common assessments that are in place across some states: the ACT Aspire, the Partnership for Assessment of Readiness for College and Careers (PARCC), and the Smarter Balanced Assessment Consortium (SBAC).[78] The ACT Aspire uses a scale that typically reports achievement in the 400-500 range (grades 3-10).[79] The PARCC scale scores range from 650 to 850 (grades 3-11).[80] The SBAC scale scores range from 2,114 to 2,795 (grades 3-8, and 11th grade).[81] Clearly, given these different scales, a scaled score cannot be compared across NAEP and a statewide assessment. Furthermore, improvement in scaled scores cannot be directly compared. If a group of students improves 20 points on the 4th grade NAEP reading assessment, it is not equivalent to a 20-point improvement on a statewide assessment, such as the PARCC or SBAC.

If scaled scores cannot be compared, what about performance standards? A performance standard is a generally agreed-upon definition of a certain level of performance in a content area that is expressed in terms of a cut score (e.g., basic, proficient, advanced) for a given assessment. There are no generally agreed-upon performance standards that apply to both NAEP and state assessments, so performance standards cannot be compared across assessments. For example, as discussed earlier, NAEP defines performance standards as basic, proficient, and advanced.[82] By contrast, PARCC and SBAC use levels as performance standards. For example, "Level 4" (out of 5) on the PARCC corresponds to "met expectations." Although it seems similar, it is unlikely that "met expectations" on the PARCC represents the same level of achievement as "proficient" on

[76] See American Institutes for Research, "New Study Examines Alignment Between NAEP and Common core State Standards in 4th, 8th Grade Mathematics," press release, October 26, 2015, https://www.air.org/news/press-release/new-study-examines-alignment-between-naep-and-common-core-state-standards-4th-8th.

[77] See, for example, the NAEP Validity Study panel report, *A Study of NAEP Reading and Writing Frameworks and Assessments in Relation to the Common Core State Standards in English Language Arts*, available at https://www.air.org/sites/default/files/downloads/report/NVS_combined_study_2_NAEP_Reading_and_Writing_Frameworks_in_Relation_to_CCSS_in_ELA_0.pdf.

[78] Although the use of common assessments has declined over recent years, they serve as a useful comparison to illustrate how statewide assessments and the NAEP can differ. For information on states' participation in common assessments, see http://educationnext.org/the-politics-of-common-core-assessments-parcc-smarter-balanced/.

[79] For more information, see https://www.discoveractaspire.org/assessments/score-scale/.

[80] For more information, see https://www.testprep-online.com/parcc-scores#scores.

[81] For more information, see https://caaspp.cde.ca.gov/sb2017/ScaleScoreRanges.

[82] *Basic* denotes partial mastery of the knowledge and skills that are fundamental for proficient work at a given grade. *Proficient* represents solid academic performance for the given grade level and competency over challenging subject matter including subject-matter knowledge, application of such knowledge to real-world situations, and analytical skills appropriate to the subject matter. *Advanced* presumes mastery of both the Basic and Proficient levels and represents superior academic performance. For more information on NAEP performance standards, see https://nces.ed.gov/nationsreportcard/tdw/analysis/describing_achiev.aspx.

the NAEP. Setting cut points for these levels requires a specific standard-setting process that is assessment-specific, so it is unlikely that meeting expectations on one assessment corresponds to the same level of performance as being proficient on another.

Perhaps even more difficult to reconcile may be when states and NAEP use the same performance standards terminology. For example, the state of Alaska uses four performance standards: Far Below Proficient, Below Proficient, Proficient, and Advanced. "Proficient" is defined as "meets the standards at a proficient level, demonstrating knowledge and skills of current grade-level content."[83] Unlike NAEP, the "proficient" definition does not necessarily include application of skills to real-world situations or analytical skills. Neither definition of "proficient" is correct or incorrect, but these definitions demonstrate the difficulty in comparing "proficient" performance standards of NAEP to those of state assessments.

In an effort to examine how closely the performance standards of NAEP reflect those used in the states, NCES released an alignment study to map state performance standards onto the NAEP scale.[84] This mapping study is not an evaluation of the quality of state performance standards or NAEP performance standards but rather is intended to give context to the discussion of comparing performance standards. The study found that most "proficient" state standards in 4th and 8th grade reading and mathematics mapped at the NAEP "basic" level. This finding reinforces the difficulty in comparing NAEP to statewide assessments. Since the "proficient" performance standard on many statewide assessments may be more comparable to the "basic" performance standard on NAEP, it may not be possible to make meaningful comparisons between state assessments and NAEP using performance standards. Given the difference in the meaning of "proficient" across assessments, the number of students "proficient" on NAEP will likely be lower than the number of students "proficient" on most state assessments. If fewer students score at the "proficient" performance standard on NAEP, it does not mean that either the NAEP or statewide assessment measured achievement correctly or incorrectly. Rather, the assessments used a different assessment framework and cut score to define the performance standard of "proficient."

NAEP and International Assessments

Both NAEP and international assessments measure reading and mathematics performance of students around 4th and 8th grade. For instance, NAEP, TIMSS, and PISA all measure mathematics performance for students around 8th grade. NAEP and PIRLS both measure 4th grade reading performance. NAEP and PISA both measure reading performance around 8th grade. Comparing NAEP to international assessments requires considering some of the same issues as comparing NAEP to statewide assessments systems: the scale, the definition of performance standards, and the alignment between the assessments. There are also some additional considerations, including the different target populations, participating education systems, differences in voluntary student participation, and the precision of measurement.

As previously discussed, NAEP and the international assessments use different scales and different performance standards to describe achievement. These types of results cannot be directly compared. Perhaps an even larger difference between NAEP and international assessments is their assessment framework, that is, the specific knowledge and skills being measured within a

[83] See Achievement Level Descriptors, available at https://education.alaska.gov/assessments/peaks.

[84] V. Bandeira de Mello, T. Rahman, and B.J. Park, *Mapping State Proficiency Standards Onto NAEP Scales: Results From the 2015 NAEP Reading and Mathematics Assessments* (NCES 2018-159), U.S. Department of Education, Washington, DC: Institute of Education Sciences, National Center for Education Statistics, 2018, https://nces.ed.gov/nationsreportcard/pubs/studies/2018159.aspx.

content area. If assessment frameworks are significantly different, the assessments are not closely aligned. While national and international assessments may appear to have significant similarities, each was designed for a different purpose and uses a unique framework to measure achievement. Differences in results across the assessments do not necessarily imply a problem with measuring achievement but rather may represent different types of achievement.

NCES has studied certain issues of alignment between NAEP, PISA, PIRLS, and TIMSS.[85] In general, NAEP was developed with national interests in mind while the international assessments were developed in a collaborative process with other countries, reflecting a consensus view of content. In terms of alignment, the NAEP and TIMSS tend to focus on "school-based learning." NAEP and TIMSS are organized similarly and measure skills such as "knowing, applying, and reasoning." In mathematics, these two assessments are relatively well-aligned, but they are less well-aligned in science.[86] PISA differs from NAEP and TIMSS in that it measures real-world learning, so it draws not only from school curricula but also learning that occurs outside of school. PISA measures mathematics skills that focus on "reproduction, connections, and reflection." In terms of reading alignment, the NAEP focuses more on school-based learning while PIRLS and PISA focus more on the context of reading and purposes for reading. There is potential overlap and potential differences in terms of the skills and abilities being assessed.[87]

In terms of target population, the NAEP, TIMSS, and PIRLS sample by grade. NAEP and TIMSS both sample students in 4th and 8th grade, which means the ages sampled are comparable. Likewise, PIRLS samples the equivalent 4th grade students, which is comparable to NAEP 4th grade students. PISA, however, samples by age. In the United States, most 15-year-old students selected for PISA are in 10th or 11th grades.[88] Comparisons to younger 8th grade students on NAEP may be less appropriate.

Different assessments have different participating education systems. An education system is typically a country but can also include a subnational jurisdiction, such as a province, state, or large city. For example, the United States participates in PISA, and Massachusetts, North Carolina, and Puerto Rico also participate as separate jurisdictions. Similarly, the United States participates in PIRLS, and Florida also participates as a separate jurisdiction. Across education systems (i.e., countries and subnational jurisdictions), there are different types of governance. Many countries have a more centralized education administration than the United States.[89] The implications for results may differ depending on the governance of the education system. For

[85] See, for example, U.S. Department of Education, National Center for Education Statistics, *Comparing TIMSS with NAEP and PISA in Mathematics and Science*, https://nces.ed.gov/timss/pdf/comparing_timss_naep_%20pisa.pdf, and U.S. Department of Education, National Center for Education Statistics, *Comparing PIRLS with PISA and NAEP in Reading, Mathematics, and Science*, https://nces.ed.gov/surveys/PISA/pdf/comppaper12082004.pdf. Also see Tom Loveless, *International Tests Are Not All the Same*, Brookings Institution, January 9, 2013, https://www.brookings.edu/research/international-tests-are-not-all-the-same/.

[86] See U.S. Department of Education, National Center for Education Statistics, *Comparing TIMSS with NAEP and PISA in Mathematics and Science*, https://nces.ed.gov/timss/pdf/comparing_timss_naep_%20pisa.pdf.

[87] See U.S. Department of Education, National Center for Education Statistics, *Comparing PIRLS with PISA and NAEP in Reading, Mathematics, and Science*, https://nces.ed.gov/surveys/PISA/pdf/comppaper12082004.pdf.

[88] See U.S. Department of Education, National Center for Education Statistics, *Comparing TIMSS with NAEP and PISA in Mathematics and Science*, https://nces.ed.gov/timss/pdf/comparing_timss_naep_%20pisa.pdf, and U.S. Department of Education, National Center for Education Statistics, *Comparing PIRLS with PISA and NAEP in Reading, Mathematics, and Science*, https://nces.ed.gov/surveys/PISA/pdf/comppaper12082004.pdf.

[89] M. Carnoy, E. Garcia, and T. Khavenson, "Bringing it Back Home: Why State Comparisons are More Useful Than International Comparisons for Improving U.S. Education Policy," Economic Policy Institute, EPI Briefing Paper #410, 2015, https://www.epi.org/publication/bringing-it-back-home-why-state-comparisons-are-more-useful-than-international-comparisons-for-improving-u-s-education-policy/.

example, while some countries can use the results of international assessments to change policies across the board, less centralized education systems may be unable to implement unilateral, system-wide changes.

International assessments also differ in how they treat the scores of subnational jurisdictions. The international average score for PISA is based only on OECD countries' scores while the international average in PIRLS is based on all participating countries and jurisdictions. Thus, across assessments and administrations, international averages are based on different sets of countries, making comparisons across time more difficult. Further complicating this issue is the fact that in each administration of an international assessment, participating countries and subnational jurisdictions can change. The average from year to year depends on which education systems participate in a particular administration.

National and international assessments remain voluntary at the individual student level. If the group of students that choose not to participate are different than students who choose to participate, it can lead to a nonrepresentative sample of participating students. If a nonrepresentative sample is assessed, it can lead to selection bias in the results.[90] On the NAEP, students with disabilities and English learners can be excluded if students require an accommodation that is not permitted by the NAEP.[91] It is possible that if many students with disabilities or ELs are excluded, the sample would not include enough of these students to be representative of the population. Typically, large-scale assessments analyze the sample for selection bias. ED provides exclusion data for students with disabilities and ELs by state for the NAEP. Exclusion rates by state vary. For example, in 2011 for 8th grade mathematics, state exclusion rates for students with disabilities and ELs ranged from 4% to 56%.[92] ED also provides exclusion rates for education systems participating in the TIMSS. For example, in 2011, education system exclusion rates ranged from 0% to 23%.[93] High exclusion rates may lead to selection bias and unreliable results that do not represent the achievement of the state or education system. Comparisons between states or education systems with high exclusion rates may be inaccurate because they may compare the achievement of a representative sample to the achievement of an unrepresentative sample.

National and international assessments differ in how precisely they can measure student achievement. The precision of measurement depends largely on the sample size. The more students that participate in an assessment, the more likely it will be to detect significant small changes in performance or performance over time. International assessments tend to sample anywhere from approximately 5,000 to 25,000 U.S. students per administration.[94] NAEP, on the

[90] Selection bias occurs when the sample selected for the assessment is not representative of the population as a whole. If selection bias is present, the results may not be applicable to the population.

[91] For specific information on the accommodations allowed on the NAEP, see https://nces.ed.gov/nationsreportcard/about/accom_table.aspx.

[92] California identified 23 students in the original sample as a student with a disability or EL and excluded 1 from the 8th grade mathematics assessment (4%). Oklahoma identified 18 students in the original sample as a student with a disability or EL and excluded 10 from the 8th grade mathematics assessment (56%). See https://nces.ed.gov/nationsreportcard/studies/naep_timss/exclusion.aspx#table1.

[93] This represents an exclusion rate for the sample overall, not specifically for students with disabilities or students who don't speak the primary language of the education system. See https://nces.ed.gov/nationsreportcard/studies/naep_timss/exclusion. aspx#table1.

[94] The procedures for selecting a sample for international assessments depends on each assessment's sampling design. The number of students sampled from each country are not equal. For example, for the 2015 PISA, the range of participating students per country ranged from approximately 3,400 students (Iceland) to 20,000 students (Canada). See Organisation for Economic Co-operation and Development, *PISA 2015 Technical Report*, Chapter 11: Sampling

other hand, samples hundreds of thousands of students.[95] NAEP, therefore, can measure student performance with more precision. Furthermore, NAEP is better suited than international assessments to measure subgroups of students due to the sampling design and overall size of the sample. Because of the differences in the precision of measurement, students may make progress on a NAEP assessment but not on an international assessment. In this case, it is possible that the international assessment did not sample enough students to detect a statistically significant increase in student performance. Similarly, NAEP results may show that an achievement gap is getting bigger or smaller, but this result may not be duplicated by international assessments. International assessments may not have a large enough sample from minority groups to detect the same size of change in achievement gap as the NAEP.

Given the difficulty of making comparisons, it is not surprising that there may be differences in results for a given year or over time among the assessments. Each assessment was developed for its own purpose and is administered in its own way, and each may present a different side of U.S. students' achievement. For example, TIMSS results may highlight how U.S. students perform on measures of school-based learning and PISA results may highlight how U.S. students perform on measures of real-world applications of learning. The fact that U.S. students perform relatively better on TIMSS than on PISA does not mean that either result is wrong but that they are measuring different skills.

Why Participate?

As previously discussed, students already participate in myriad assessments at the state and local levels, including state assessments in reading, mathematics, and science required under ESEA Title I-A. These state assessments must be aligned with state standards in the relevant subject areas. From a policy perspective, this raises obvious questions about why the United States participates in NAEP and international large-scale assessments when data on student performance are available from a multitude of other assessments.

NAEP Participation

While every state administers state assessments aligned with state standards in reading, mathematics, and science, each state is able to select its own assessment and its own standards. Thus, it is possible that every state could use a different set of assessments aligned with different content and performance standards, making it difficult to compare student achievement across states based on these data. NAEP is a nationally administered assessment in reading and mathematics, and periodically in other subjects, which produces nationally representative data for each state and for large urban districts. It provides a comparison of student achievement across the United States in the subjects tested.[96] It also provides both a snapshot in time of student achievement and trends in this achievement over time for the nation overall, states, and large

Outcomes, 2017, pp. 203-250, http://www.oecd.org/pisa/sitedocument/PISA-2015-technical-report-final.pdf.

[95] For example, in the 2017 administration of the main NAEP, approximately 585,000 students participated.

[96] For more information about the NAEP, see U.S. Department of Education, National Center for Education Statistics, *An Introduction to NAEP*, 2010, https://nces.ed.gov/nationsreportcard/about/.

urban districts, and allows comparisons of student subgroup performance at each of these levels.[97] NAEP enables states to benchmark state performance against other states and the nation.[98]

According to NAEP, the data are also used to "inform educational policy and practice by

- reporting the achievement of various student groups,
- analyzing NAEP results in the context of educational experiences, and
- providing tools and resources for data analysis."[99]

International Assessment Participation

NAEP provides data on national, state, and large urban district performance. By itself, however, it does not provide any data on the how students in the United States compare to those in other countries. Participation in international large-scale assessments provides insights into how the United States performs relative to other countries based on external standards that are measured the same way for each country participating in a given assessment.[100]

These assessments can provide another piece of information that broadly addresses student achievement and academic trends in the United States, potentially confirming or contradicting evidence from other assessments. Other reasons cited in research for participating in international large-scale assessments include being able to benchmark state standards to those of other countries; examine educational progress over time among countries (as opposed to only states in the United States); learn more about what is educationally possible to establish performance expectations; collect information about school environments, instruction, and resources; and compare the performance of groups of students (e.g., by race/ethnicity) with comparable groups of students from other countries to examine achievement gaps and other issues.[101] Some researchers have also argued that the data from other countries can provide a "unique basis for generating hypotheses about American secondary schooling," even when the education systems of other countries are considerably different.[102] If a country scores well on an international assessment, researchers may be able to isolate policies and practices that may have contributed to

[97] U.S. Department of Education, National Center for Education Statistics, *NAEP Data Informs Policy and Practice*, https://nces.ed.gov/nationsreportcard/about/policy_practice.aspx.

[98] See, for example, Valena White Plisko, "Participation in International Large-Scale Assessments from US Perspective," *Research in Comparative and International Education*, vol. 8, no. 3 (2013).

[99] U.S. Department of Education, National Center for Education Statistics, *NAEP Data Informs Policy and Practice*, https://nces.ed.gov/nationsreportcard/about/policy_practice.aspx.

[100] Valena White Plisko, "Participation in International Large-Scale Assessments from US Perspective," *Research in Comparative and International Education*, vol. 8, no. 3 (2013).

[101] See, for example, Gary W. Phillips, *International Benchmarking: State and National Performance Standards*, American Institutes for Research, September 2014, https://www.air.org/sites/default/files/downloads/report/AIR_International%20Benchmarking-State%20and%20National%20Ed%20Performance%20Standards_Sept2014.pdf; David J. Rutkowski and Ellen L. Prusinski, "The Limits and Possibilities of International Large-Scale Assessments," *Center for Evaluation and Education Policy: Education Policy Brief*, vol. 9, no. 2 (Spring 2011); and Valena White Plisko, "Participation in International Large-Scale Assessments from US Perspective," *Research in Comparative and International Education*, vol. 8, no. 3 (2013).

[102] For example, while Daniel Koretz acknowledges that the United States may differ in many ways from high scoring countries (e.g., system of governance, instructional methods) and the assessments cannot pinpoint which factors may contribute to the differences in scoring, he argues that they do provide suggestions that can be tested with appropriate study designs. See Daniel Koretz, "How Do American Students Measure Up? Making Sense of International Comparisons," *Future child*, vol. 19, no. 1 (Spring 2009), p. 48. See also, for example, Valena White Plisko, "Participation in International Large-Scale Assessments from US Perspective," *Research in Comparative and International Education*, vol. 8, no. 3 (2013).

the country scoring well. Researchers can select these policies and practices and develop hypotheses about how they may affect the achievement of students in the United States. These policies and practices, however, are not de facto effective practices that can be immediately implemented in the United States, but rather, require further study within the context of the U.S. education system.

Limitations of NAEP and International Large-scale Assessments for Policy Consideration

While there are several reasons that the United States chooses to administer NAEP and to participate in international large-scale assessments, there are several factors that limit the use of national and international assessment results in shaping education policy. In addition, it is unclear whether student achievement on international assessments may be related to economic prosperity and whether increasing student achievement on these assessments may be linked to improvements in a country's economic health.

Identification and Implementation of Policies to Increase Achievement on the Basis of National and International Assessments

When national and international assessment results are released, they provide a snapshot of the general condition of education. Tracking results over time can indicate whether students are making educational progress in certain content areas at certain grades. If U.S. students are ranked significantly lower than many other countries and not making clear progress over time, it may signal a problem in elementary and secondary education policies and practices. The results, however, may not be particularly helpful in identifying policies that may increase student achievement or aid the United States in meeting other educational goals, such as increasing high school graduation rates. For example, U.S. students' achievement has significantly decreased over the last two administrations of PISA. A decrease in achievement could represent an actual decrease in student achievement. On the other hand, a decrease in achievement could be indicative of curricula that are misaligned with the test, teaching and learning practices in the United States that are different than what PISA requires, or even a lack of student engagement in the testing process.[103] With the numerous possibilities to explain student achievement, it may be unclear how policymakers should begin to address a decrease in achievement.

One way policymakers may consider addressing a decrease in achievement is to adopt policies and practices from countries that consistently score well on international assessments. This approach raises myriad questions. For example, is it in the best interest of the United States and its students to adopt the educational policies of countries that may be quite different than the United States? Other countries may differ from the United States in many ways, including with respect to their student populations, levels and distribution of education funding, policies regarding the tracking of students by academic ability, secondary school and university enrollment rates, and the quality of educators provided to subgroups of students (e.g., low-income students, English learners, students with disabilities, minority students).[104] And if it is determined

[103] For more information, see Stephen Sawchuk, "If Students Aren't Trying on International Tests, Can We Still Compare Countries' Results?" *Education Week*, August 22, 2018, https://www.edweek.org/ew/articles/2018/08/22/if-students-arent-trying-on-international-tests.html.

[104] Iris Rotberg, "Assessment Around the World," *Educational Leadership*, vol. 64, no. 3 (November 2006), pp. 58-63.

that adopting the policies of another country is the best course of action, the feasibility of adopting such policies must be addressed, including whether educational policies are generalizable across countries and whether the will and capacity to make needed changes exists. In addition, just as academic achievement in the United States has changed over time, the same is true for other countries. For example, earlier in the 21st century, Finland was a top performer on PISA and viewed as a country having policies worthy of emulation.[105] However, in both the 2012 and 2015 administrations of the PISA, Finland has seen its performance in science, math, and reading decline. This raises questions about what would have happened if the United States had focused on mirroring Finland's policies in the hopes of achieving Finland's top level of performance.[106]

The example of Finland above highlights an important characteristic of international assessments—these assessments provide a snapshot of achievement, but they do not evaluate policies and practices within or across countries. The PISA did not evaluate the effectiveness of any policies or practices in place in Finland. The decline in achievement seen in Finland in the 2012 and 2015 PISA assessments may have been due to a specific policy or practice or it may have been due to factors outside of education (e.g., economic health, political climate, changing demographics, etc.). In the 1990s and early 2000s, PISA was not able to provide information on why students in Finland were high achieving, and currently, PISA is not able to provide information on why the achievement of students in Finland has declined. None of the international assessments provides data on the factors that may explain student achievement.

Even if it is possible to identify an education policy that would increase U.S. student achievement on national and international assessments, there may be barriers to implementing that policy. Compared to many other countries that participate in international assessments, the United States has a decentralized education system that primarily reserves the power to make education policy decisions for state and local authorities. State and local authorities already rely upon state and local assessments to evaluate students, schools, and districts. While the results of national and international assessments may, in some cases, highlight a problem, the assessment results may not offer a policy solution for states. By contrast, state and local assessments are aligned with state content and performance standards, which possibly make them better suited than national and international assessments to address any perceived problems in teaching and learning at the state level. For example, if student performance is trending downward in reading on a statewide assessment, state and local authorities may choose new curricula, allocate more teaching time to reading, or provide funding for reading specialists. If student performance is trending downward in reading on an international assessment, state and local authorities cannot determine whether the score is low because student achievement is actually declining or if the test is not aligned with their content standards, curricula, or teaching practices.

Impact of Achievement on Economic Prosperity

There is considerable debate about the impact of student achievement on a country's economic prosperity. Several analyses have tried to link performance on international assessments to various indicators of national wealth and prosperity. One analysis of the relationship between

[105] For more information about Finland's performance on NAEP, see Joe Heim, "Finland's schools were once the envy of the world. Now, they're slipping," *Washington Post*, December 8, 2016, available online at https://www.washingtonpost.com/local/education/finlands-schools-were-once-the-envy-of-the-world-now-theyre-slipping/2016/12/08/dcfd0f56-bd60-11e6-91ee-1adddfe36cbe_story.html?utm_term=.c9175ae2cb6d.

[106] It should be noted that despite declines in achievement, Finland continues to outperform the United States on PISA (ibid.).

international test scores and "national success" that garnered attention was presented in 2007 by Keith Baker, a former researcher at ED.[107] Baker used scores from 11 countries that participated in the First International Mathematics Study (FIMS) in 1964 to predict seven indicators of national success 30 years later: wealth, rate of growth, productivity, quality of life, livability, democracy, and creativity. For almost all indicators, there was no relationship or a negative relationship between scores on FIMS and national success. That is, as scores on international assessments increased, the wealth, rate of growth, etc. of a nation decreased or remained the same.[108] Increases in achievement, therefore, were associated with decreases in indicators of national success. The one exception was the indicator of creativity (as measured by the number of patents issued). As international test scores increased, the creativity indicator also increased.

Baker also analyzed the relationship between PISA and national success indicators. Results showed that nations at the PISA average generally outperformed other nations scoring well above or well below average. Based on these findings, Baker concluded that there may be some baseline level of achievement that is important for national success; however, once that baseline has been reached, focusing on increasing test scores may divert time and resources away from other factors that may contribute to national success.

Another analysis that garnered attention was conducted by Eric Hanushek and Ludgar Woessmann and published by the OECD.[109] Their analysis used economic modeling to predict the impact of achievement on international assessments on economic growth. The model used growth in PISA scores over time to project growth in gross domestic product (GDP) in selected countries. The results indicate that if all OECD countries increased their average PISA scores by 25 points over the next 20 years, the aggregate gain of OECD GDP over 80 years would be approximately $115 trillion. Furthermore, if all countries performed at a level of minimal proficiency for the OECD,[110] the aggregate GDP would increase $200 trillion.[111]

Without a clearer picture of how performance on international assessments contributes to a country's economic prosperity, it may be difficult to decide whether attempting to increase international test scores for this purpose would be a worthwhile education policy goal. Given

[107] Keith Baker, "Are International Tests Worth Anything?" *Phi Delta Kappan*, vol. 89, no. 2 (October 2007), pp. 101-104.

[108] There are some drawbacks to consider from the analysis of FIMS data. First, this was a retrospective, correlational analysis that does not imply causality. Since no causal relationship can be established with correlations, it is impossible to know whether an increase in international test scores necessarily leads to a decrease in national success indicators, as the correlation may suggest. Second, the early administrations of international assessments had fewer participating countries, which may limit the generalizability of these findings. For example, FIMS was administered in 11 countries in 1964. In 2015, TIMSS (which is considered the current iteration of FIMS) is administered in more than 60 education systems.

[109] Eric Hanushek and Ludgar Woessmann, *The High Cost of Low Educational Performance: The Long-Run Impact of Improving PISA Outcomes*, Organisation for Economic Co-operation and Development, 2010, https://www.oecd.org/pisa/44417824.pdf.

[110] The researchers defined "minimal proficiency" as a PISA score of 400.

[111] There are some drawbacks to consider when using economic modeling with achievement indicators, such as scores on PISA. First, as with the Baker analysis described earlier, the model does not prove causality. Second, the model assumes a linear relationship between the PISA and GDP, which may not hold across all possible scores on the PISA. That is, there may be a certain level of achievement at which no further gains in GDP would be seen. It is possible that there is some level of achievement at which further gains in GDP would no longer be observed. For example, in the Baker analysis, he concluded that there may be a certain baseline level of achievement important for national success; however, above this baseline, further gains in economic prosperity may not be seen. If this is the case, the relationship between international assessment scores and economic prosperity is not linear.

limited time and resources, policymakers may choose to focus efforts on other factors that contribute to educational achievement and economic prosperity.

Concluding Thoughts

Given the overwhelming amount of data gathered by national and international assessments, it is difficult to comb through all of the results and gain a clear picture of the achievement of U.S. students over time and relative to other countries. Perhaps even more difficult is understanding the policies and practices that drive performance on these assessments. For example, in the last decade, two Secretaries of Education have expressed concern over U.S. performance on the PISA and called for different reforms to address the problem. In 2010, Secretary Arne Duncan used U.S. performance on the PISA to argue for advancing the education policy goals of the Obama Administration, most notably changes to teacher recruitment, teacher evaluation, and the compensation of highly effective teachers. He argued that the OECD found that most high-achieving countries in PISA have policies that mirror the Obama Administration's focus on highly effective teachers.[112] By contrast, Secretary Betsy DeVos used U.S. performance on PISA to argue for advancing the education policy goals of the Trump Administration, most notably school choice. She argued that countries that outperform the United States on PISA have more quickly adopted school choice policies.[113] Clearly, education leaders have been concerned about the performance of U.S. students on international assessments; however, the data from the assessments do not point to either a conclusive policy problem or solution.

Both NAEP and international large-scale assessments provide the United States with comparative data about student achievement that is not available through assessments administered only at the state and local level. Having these data to examine student performance at a given moment in time and over the long term can be used in many ways, including as a check to confirm or contradict what data from state and local assessments indicate about student performance.

While NAEP is developed and implemented solely in the United States, the results of the assessments may have limited value in identifying particular policies or practices that are and are not working for U.S. students. Since NAEP is not aligned with any particular state's standards or curricula, it cannot provide direct feedback on how well students within a state are achieving the state's standards. Although, NAEP results do provide states with an opportunity to benchmark themselves against other states that operate in the same decentralized system of educational control. International large-scale assessments also offer benchmarking and like NAEP may be less useful with respect to determining which policies and practices are contributing to student success and whether those policies and practices could be successfully implemented in locales within the United States. Thus, while national and international assessments may provide valuable data, the data do not easily translate into effective education policies. The results, therefore, may be more useful in identifying areas in need of attention or resources and may have limited utility for shaping education policy approaches.

---

[112] See Secretary Arne Duncan's remarks at the OECD's Release of the Program for International Student Assessment (PISA) 2009 Results, U.S. Department of Education, *Secretary Arne Duncan's Remarks at OECD's Release of the Program for International Student Assessment (PISA) 2009 Results*, December 7, 2010, https://www.ed.gov/news/speeches/secretary-arne-duncans-remarks-oecds-release-program-international-student-assessment-pisa-2009-results.

[113] See, for example, Alyson Klein, "Betsy DeVos Links Nation's Stagnant Test Scores to Lack of Parental Choice," *Education Week*, November 30, 2017, http://blogs.edweek.org/edweek/campaign-k-12/2017/11/betsy_devos_test_scores_stagnant_parental_choice.html, and U.S. Department of Education, *Prepared Remarks by U.S. Education Secretary Betsy DeVos to the American Enterprise Institute*, January 16, 2018, https://www.ed.gov/news/speeches/prepared-remarks-us-education-secretary-betsy-devos-american-enterprise-institute.

Appendix A. National and International Educational Assessments: Authorization and Oversight Provisions

Table A-1. Large-Scale Assessment Authorization and Oversight

Assessment Title	Authorization	Oversight
National Assessment of Educational Progress (NAEP)	National Assessment of Educational Progress Assessment Act (NAEPAA; Title III, Section 303 of P.L. 107-279).	The Commissioner of the National Center for Education Statistics (NCES) at the U.S. Department of Education (ED) administers NAEP. The National Assessment Governing Board (NAGB) sets the policy for NAEP.[a]
Trends in International Mathematics and Science Study (TIMSS)	International assessment activities are authorized by the Education Sciences Reform Act (ESRA; P.L. 107-279, Section 153(a)(6)).	Assessments for U.S. students are organized under the International Activities Program and administered by NCES.
Program for International Student Assessment (PISA)	International assessment activities are authorized by the ESRA (P.L. 107-279, Section 153(a)(6)).	Assessments for U.S. students are organized under the International Activities Program and administered by NCES.
Progress in International Reading Literacy Study (PIRLS)	International assessment activities are authorized by the ESRA (P.L. 107-279, Section 153(a)(6)).	Assessments for U.S. students are organized under the International Activities Program and administered by NCES.

Source: CRS summary of national and international assessments, available from the U.S. Department of Education (ED).

a. NAGB was created in 1988 and is currently authorized under NAEPAA, Section 302. NAGB's primary policy document, *General Policy: Conducting and Reporting the National Assessment of Educational Progress*, can be found at https://www.nagb.gov/content/nagb/assets/documents/policies/GP-Conducting-and-Reporting-National-Assessment-of-Educational-Progress.pdf.

Appendix B. Additional Resources on National and International Assessments

For more information on NAEP:

- https://nces.ed.gov/nationsreportcard/

For more information on TIMSS:

- https://nces.ed.gov/timss/
- https://iea.nl/timss/

For more information on PIRLS:

- https://nces.ed.gov/surveys/pirls/
- https://iea.nl/pirls/

For more information on PISA:

- https://nces.ed.gov/surveys/pisa/
- http://www.oecd.org/pisa/

For more information on the coordination of NAEP and international assessments:

- https://nces.ed.gov/nationsreportcard/about/international.aspx

Other CRS reports on assessment in elementary and secondary education:

- CRS In Focus IF11021, *National and International Educational Assessments*
- CRS Report R45048, *Basic Concepts and Technical Considerations in Educational Assessment: A Primer*
- CRS Report R45049, *Educational Assessment and the Elementary and Secondary Education Act*

Appendix C. Glossary of Acronyms

CCSS: Common Core State Standards

ED: U.S. Department of Education

EL: English learner

ESEA: Elementary and Secondary Education Act

ESRA: Education Sciences Reform Act (P.L. 107-279, Title I)

ESSA: Every Student Succeeds Act (P.L. 114-95)

FIMS: First International Math Study

GDP: Gross domestic product

IEA: International Association for the Evaluation of Educational Achievement

LTT NAEP: Long-term trends National Assessment of Educational Progress

NAEP: National Assessment of Educational Progress

NAEPAA: National Assessment of Educational Progress Assessment Act (P.L. 107-279, Title III)

NAGB: National Assessment Governing Board

NCES: National Center for Education Statistics

NCLB: No Child Left Behind Act (P.L. 107-110)

NSLP: National School Lunch Program

OECD: Organisation for Economic Co-operation and Development

PARCC: Partnership for Assessment of Readiness for College and Careers

PIRLS: Progress in International Reading Literacy Study

PISA: Program for International Student Assessment

SBAC: Smarter Balanced Assessment Consortium

TUDA: Trial Urban District Assessment (part of NAEP)

TIMSS: Trends in International Mathematics and Science Study

Author Information

Rebecca R. Skinner
Specialist in Education Policy

Acknowledgments

Erin Lomax, former CRS analyst and current independent contractor to CRS, co-authored this report.

Disclaimer